ADVANCED BIOMEDICAL COMPUTING CENTER

# Annotation, Visualization, and Impact Analysis (AVIA) Tutorial

v. 1.0

**2/6/2013**

A plethora of information that emerges from large-scale genome characterization studies has triggered a development of computational frameworks and tools for efficient analysis, interpretation and visualization of genomic data.  Functional annotation of genomic variations and the ability to visualize the data in the context of whole genome and/or multiple genomes has remained a challenging task. We have developed an interactive web-based tool, AVIA (Annotation, Variation and Impact Analysis), to explore and interpret large sets of genomic variations (single nucleotide variations (SNVs) and insertion/deletions) to help guide and summarize genomic experiments.

Table of Contents

# AVIA Tutorial

## I) What is AVIA?

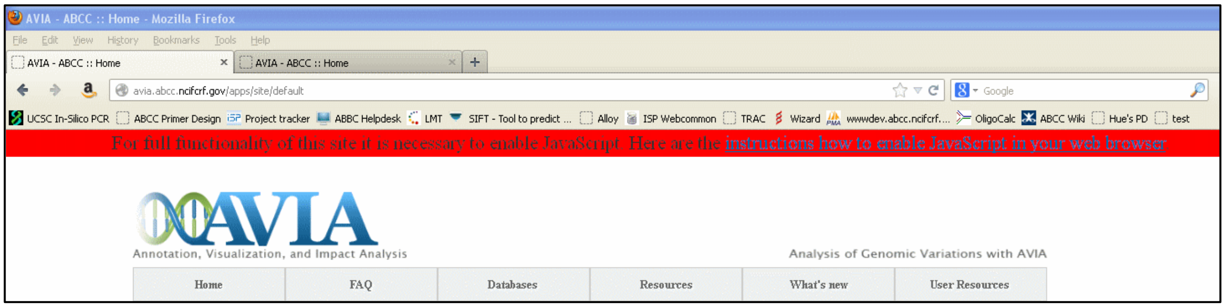
AVIA, or Annotation, Variation and Impact Analysis, is a web server dedicated to annotation of genomic variations ( SNPs and InDels) found through the high-throughput sequencing. It utilizes ANNOVAR (www.openbioinformatics.org/annovar/) as the core computational framework for assigning functional impact to genomic variations. AVIA aggregates a variety of annotation databases and analysis tools for comprehensive analysis and visualization of variation data. The annotation in AVIA is gene-centric. The default source of the basic gene set is RefSeq annotation; however, Ensembl may be selected upon request.

## II) What do I need to get started?

There are several workflows available through AVIA. Each workflow accepts mapped genomic variation data in one of the formats: VCF4, BED, CLC Bio, and HGVS. For training purposes of this tutorial, we will use sample data from the Riken Liver Cancer set from http://www.icgc.org.

You will also need an Internet web browser (not Microsoft Internet Explorer) with javascript enabled. You will not be able to submit to AVIA unless javascript is enabled. If you do not have javascript enabled, you will see the error message at the top of the page, as shown in Figure 1.

**Figure 1. Javascript Not Enabled**



For full functionality of this site it is necessary to enable JavaScript. Here are the instructions how to enable JavaScript in your web browser

As shown in the banner, if you do not know how to enable javascript, please see this tutorial at http://enable-javascript.com/. Be sure to close your browser after enabling javascript or reload the AVIA web page so that the changes can take effect.

## III) File Formats

Like ANNOVAR, the preferred input format is a tab delimited **BED**-like file (http://genome.ucsc.edu/FAQ/FAQformat.html#format1). All other formats will be converted into BED-like file and used for analysis for consistency throughout the pipeline. The BED-like format has five required fields: chromosome, genomic start, genomic stop, reference allele and variant allele. All fields after the fifth column are considered as comments and are ignored by the software, but included in the final report. A "." (dot) may be substituted for the reference allele. All variants must be the same orientation (+/+) of the genome. If headers are included, please add one line to the beginning of the text file beginning with "#". For example:

| #Chr | Start | Stop | Ref | Allele | SampleInfo … |
|---|---|---|---|---|---|

The program will report an "ERR" for annotations on any lines that do not start with the 5 required fields or does not start with "#".

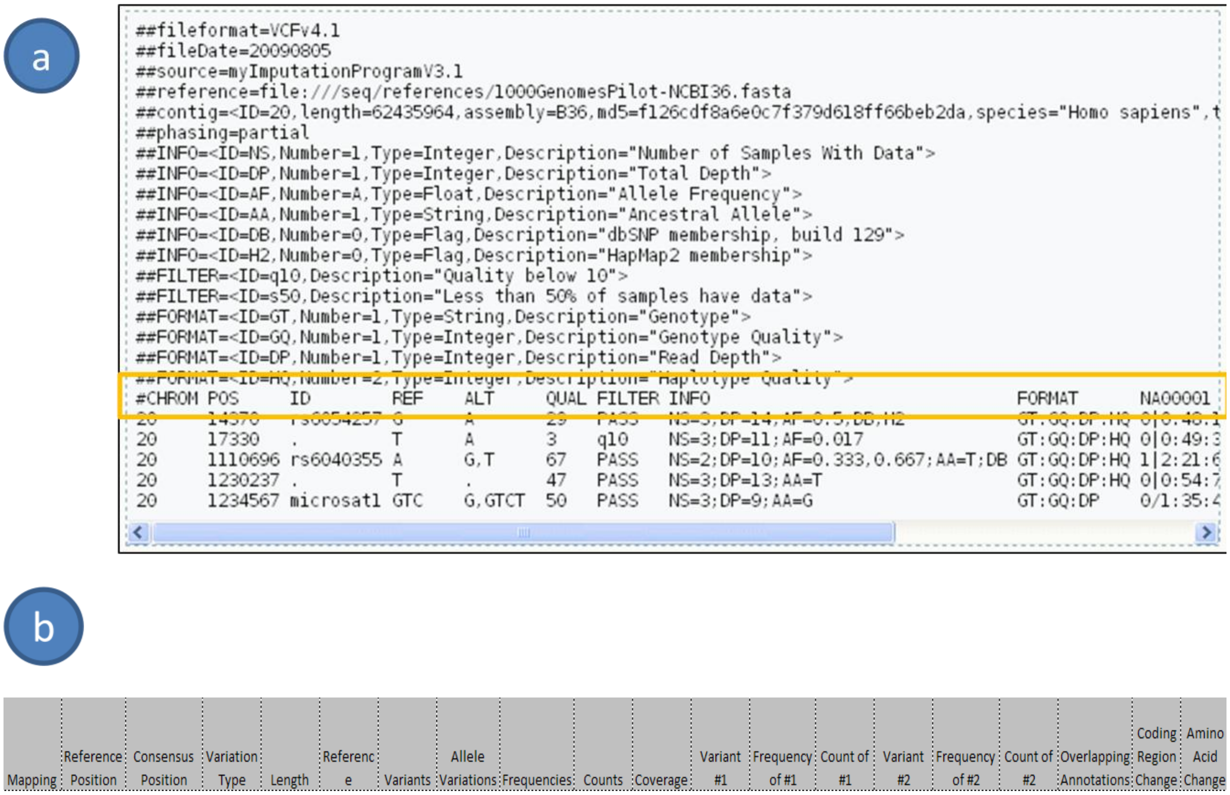
**VCF**, or variant call format, as described by University of California Santa Cruz (UCSC), "is a flexible and extendable format for variation data such as single nucleotide variants, insertions/deletions, copy number variants and structural variants." **VCF** is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines, each containing information about a position in the genome.

The following data types also accepted to AVIA, but are converted to the BED-like format described above.

1. **VCF4** format: An example of the **VCF** format from the 1000Genomes page is shown in Figure 2a. (http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41). Comment lines beginning with **"##" a**re ignored; however, it is necessary to have the final header, beginning with the "#CHROM" to parse the data correctly shown in Figure 2a, orange box.
2. **CLCBio**: The header line from **CLC Bio's** proprietary genomics workbench output is shown in Figure 2b**.**

3. **HGVS:** The final accepted format standard is the Human Genome Variation Society (HGVS) for DNA nomenclature and is provided at http://www.hgvs.org/mutnomen/recs-DNA.html.

**Figure 2. Other File Formats**

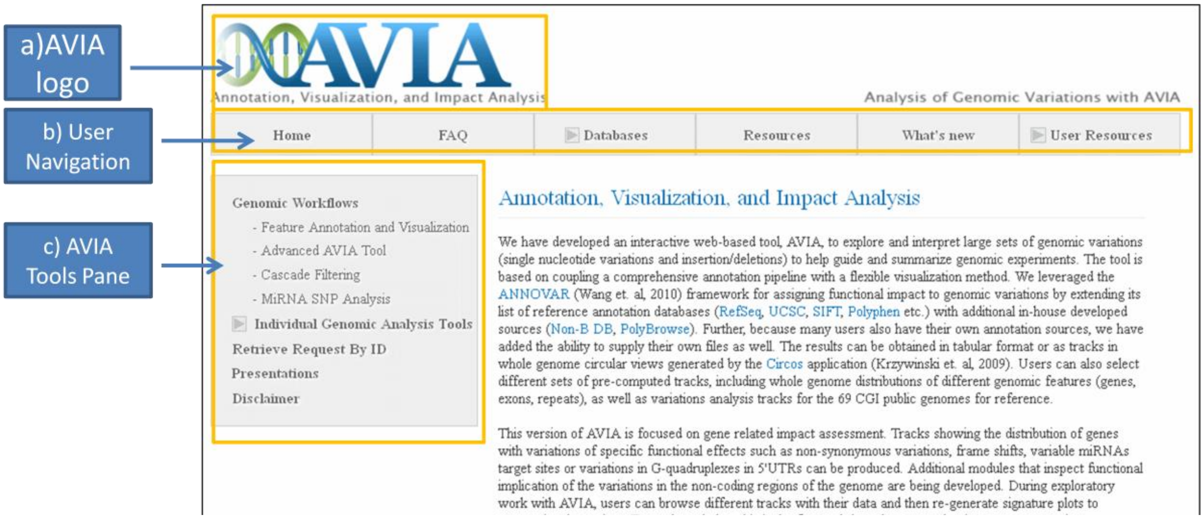


```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",t
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

| Mapping | Reference Position | Consensus Position | Variation Type | Length | Referenc e | Variants | Allele Variations | Frequencies | Counts | Coverage | Variant #1 | Frequency of #1 | Count of #1 | Variant #2 | Frequency of #2 | Count of #2 | Overlapping Annotations | Coding Region Change | Amino Acid Change |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

## IV) Basic Navigation of the Site

**Figure 3. Navigation Basics**



Annotation, Visualization, and Impact Analysis

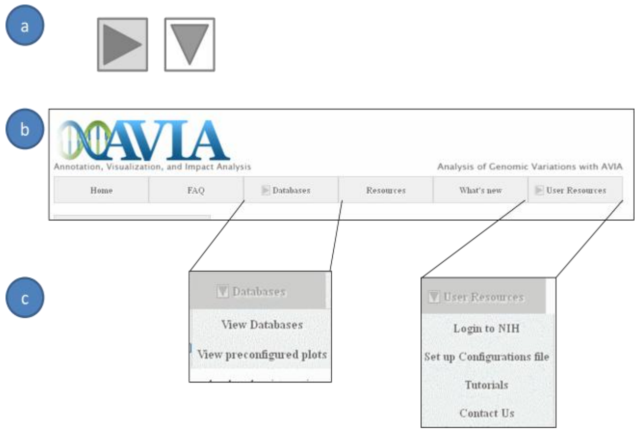
We have developed an interactive web-based tool, AVIA, to explore and interpret large sets of genomic variations (single nucleotide variations and insertion/deletions) to help guide and summarize genomic experiments. The tool is based on coupling a comprehensive annotation pipeline with a flexible visualization method. We leveraged the ANNOVAR (Wang et. al, 2010) framework for assigning functional impact to genomic variations by extending its list of reference annotation databases (RefSeq, UCSC, SIFT, Polyphen etc.) with additional in-house developed sources (Non-B DB, PolyBrowse). Further, because many users also have their own annotation sources, we have added the ability to supply their own files as well. The results can be obtained in tabular format or as tracks in whole genome circular views generated by the Circos application (Krzywinski et. al, 2009). Users can also select different sets of pre-computed tracks, including whole genome distributions of different genomic features (genes, exons, repeats), as well as variations analysis tracks for the 69 CGI public genomes for reference.

This version of AVIA is focused on gene related impact assessment. Tracks showing the distribution of genes with variations of specific functional effects such as non-synonymous variations, frame shifts, variable miRNAs target sites or variations in G-quadruplexes in 5'UTRs can be produced. Additional modules that inspect functional implication of the variations in the non-coding regions of the genome are being developed. During exploratory work with AVIA, users can browse different tracks with their data and then re-generate signature plots to

The AVIA default page is shown in Figure 3. You can always return to this page by clicking on the "AVIA" logo at the top left corner (Figure 3a) or on the Home tab (Figure 3b, first tab). Along the top, outlined in orange is the “User Navigation” Panel (Figure 3b). These tabs are user specific tools and are helpful to find useful information about our site. On the left side of the page is the AVIA Tools panel (Figure 3c). This is where the users will choose options to run the AVIA workflows using their data. We will get to those in the next section under Section IV Workflows.

From Figure 4a, you can see the two arrows that we are using to indicate whether a section of the website is expandable or collapsible. The dark gray right arrow indicates that the header may be clicked on and more options will appear. The dark gray down arrow against the white background indicates that the section can be collapsed, hiding some options. We will now briefly describe the User Navigation Panel, as this is where you will find many answers to your questions about the pipeline. The User Navigation Panel is shown in Figure 4b. You will notice that both “Databases” and “User Resources” tab have the right arrow icon. If you hover over that tab with your mouse, more options will appear, as seen in Figure 4c. Below, we briefly describe each tab and its sub categories.

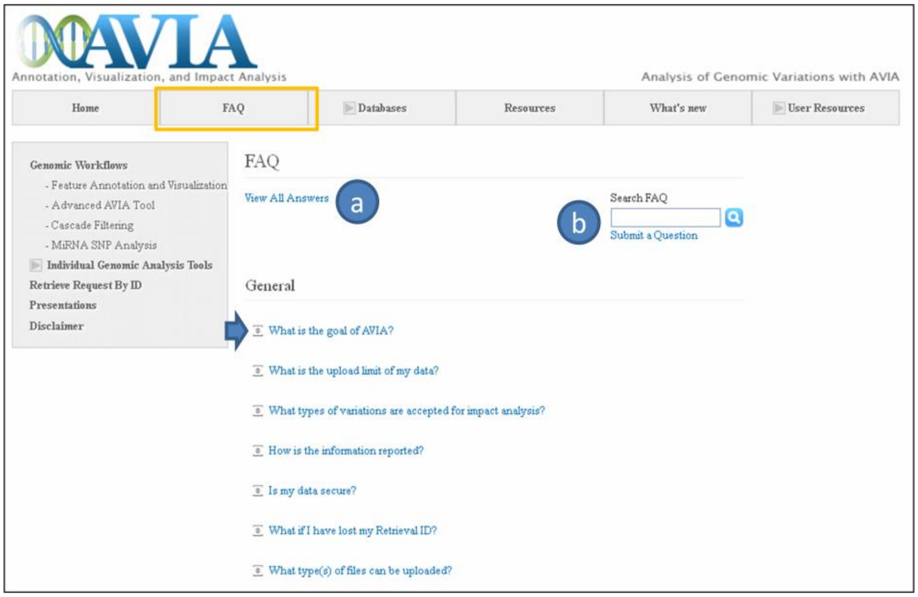
**Figure 4. Additional Options**



## A) FAQ

By clicking on the “FAQ” tab on the User Navigation Panel, it will bring you to the FAQ section of AVIA. From this page you will see our frequently asked questions regarding the pipeline and output.

**Figure 5. FAQ Section**



What is the goal of AVIA?

What is the upload limit of my data?

What types of variations are accepted for impact analysis?

How is the information reported?

Is my data secure?

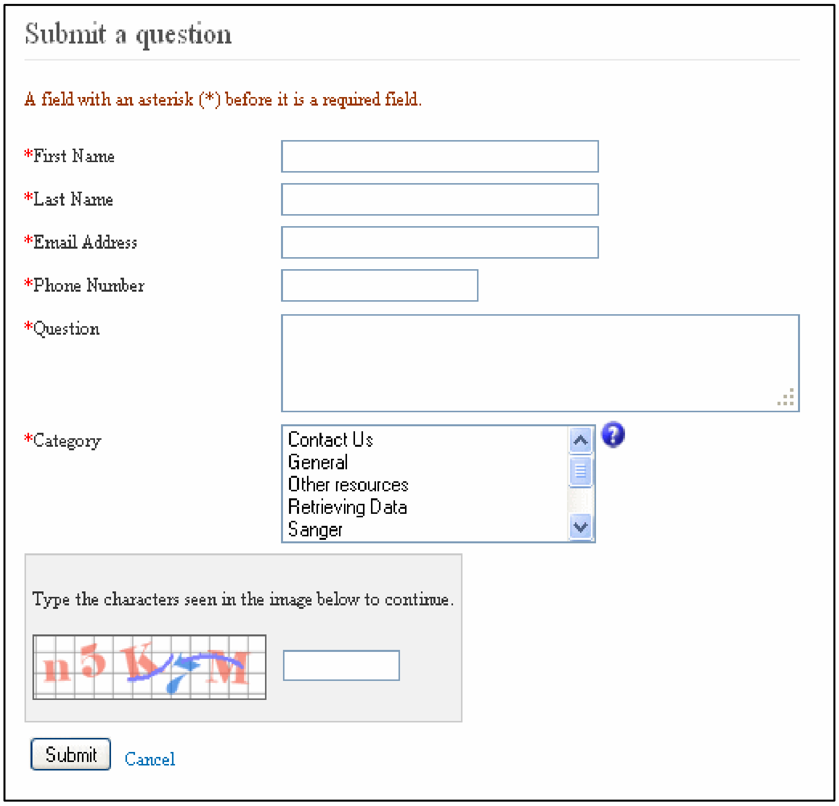
What if I have lost my Retrieval ID?

What type(s) of files can be uploaded?

You can choose to (Figure 5a) expand all answers to view all the answers or click on an individual question (Figure 5, blue arrow) to view the answers.  You may also use the Search feature (Figure 5b) to look for a specific topic.

If you cannot find the topic that you are interested in using the Search feature, you can also click on "Submit a Question" below the search field.  Using our web form shown in Figure 6, you can submit using all of the required fields.  An AVIA team member will respond to your question or comment as soon as possible.

**Figure 6. Submit A Question**



Submit a question

A field with an asterisk (*) before it is a required field.

Type the characters seen in the image below to continue.

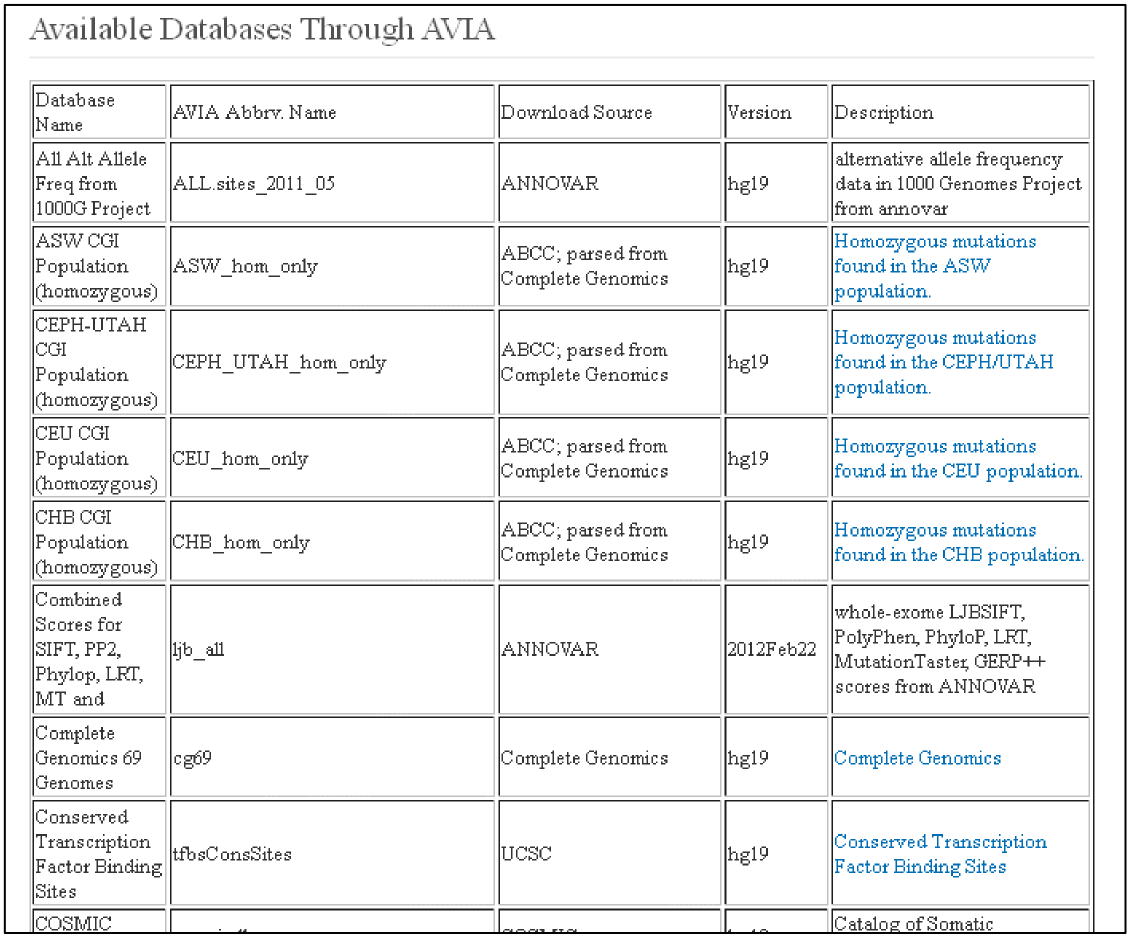
## Databases

Under the “Databases” tab, there are two options: “View Databases” and “View Preconfigured Plots” as shown in Figure 4c.

### 1) Database Information

In the “View Databases” Option, you will be able to view the most recent compiled collection of annotation databases, AVIA’s Abbreviated name, the source from which data was obtained, the version or the download date and a brief description of the database with links, if available. The databases shown in Figure 7 are a snippet and may differ from the current version. Also, when you submit to AVIA, all of the annotation is available in your total archive file, which we will discuss in VII “Output Interpretation” section of the tutorial.

**Figure 7. Available Databases**



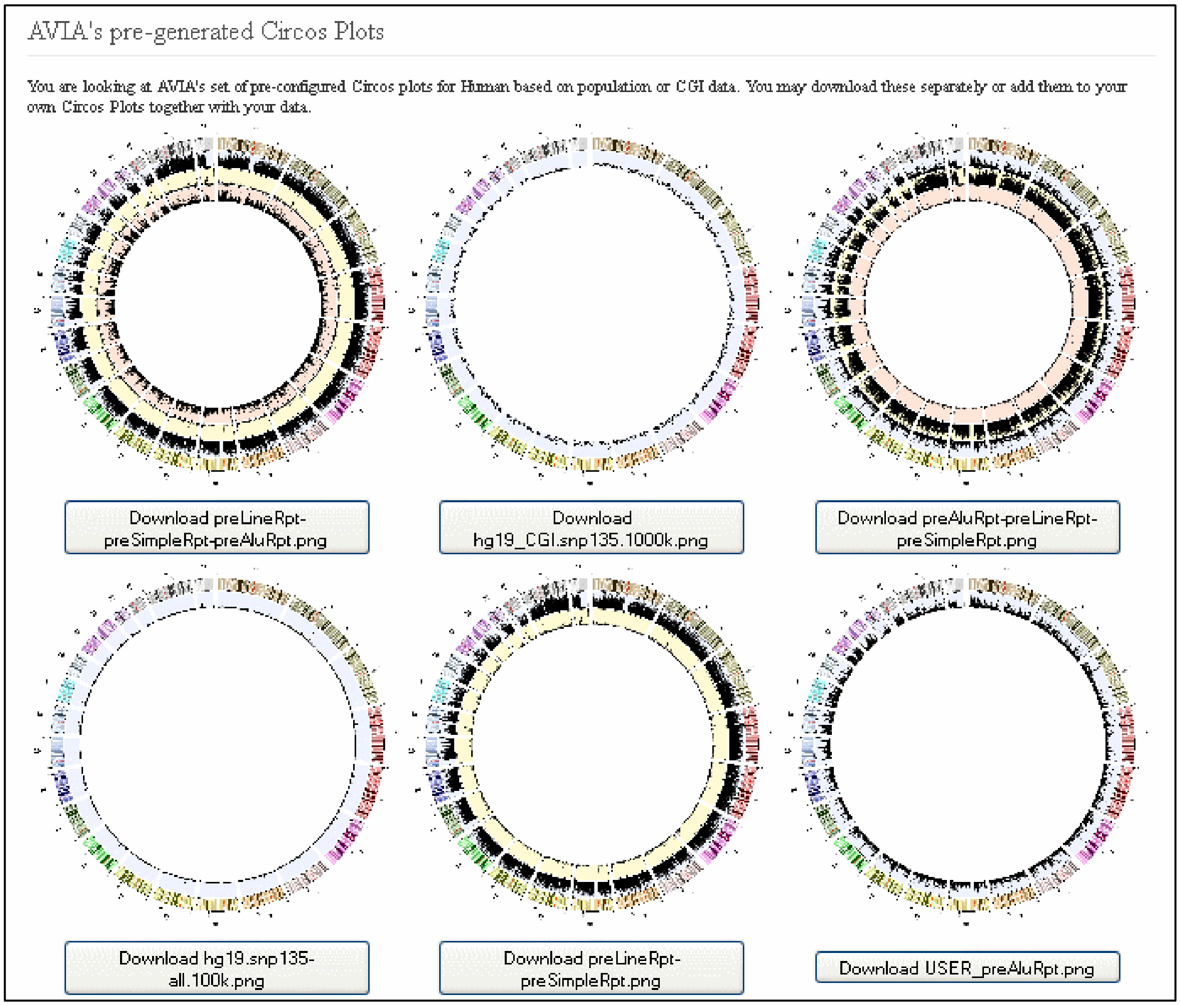
Available Databases Through AVIA

| Database Name | AVIA Abbrv. Name | Download Source | Version | Description |
|---|---|---|---|---|
| All Alt Allele Freq from 1000G Project | ALL.sites_2011_05 | ANNOVAR | hg19 | alternative allele frequency data in 1000 Genomes Project from annovar |
| ASW CGI Population (homozygous) | ASW_hom_only | ABCC; parsed from Complete Genomics | hg19 | Homozygous mutations found in the ASW population. |
| CEPH-UTAH CGI Population (homozygous) | CEPH_UTAH_hom_only | ABCC; parsed from Complete Genomics | hg19 | Homozygous mutations found in the CEPH/UTAH population. |
| CEU CGI Population (homozygous) | CEU_hom_only | ABCC; parsed from Complete Genomics | hg19 | Homozygous mutations found in the CEU population. |
| CHB CGI Population (homozygous) | CHB_hom_only | ABCC; parsed from Complete Genomics | hg19 | Homozygous mutations found in the CHB population. |
| Combined Scores for SIFT, PP2, Phylop, LRT, MT and | ljb_all | ANNOVAR | 2012Feb22 | whole-exome LJBSIFT, PolyPhen, PhyloP, LRT, MutationTaster, GERP++ scores from ANNOVAR |
| Complete Genomics 69 Genomes | cg69 | Complete Genomics | hg19 | Complete Genomics |
| Conserved Transcription Factor Binding Sites | tfbsConsSites | UCSC | hg19 | Conserved Transcription Factor Binding Sites |
| COSMIC | | | | Catalog of Somatic |

### 2) View Pre-Configured Circos Plots

For visualization, we have pre-computed several Circos Plots based on variants found whole databases or populations of data. For example, we plotted all simple repeats in one Circos plots, or all SNPs found in CGI variant data (see Figure 8; first two Circos plots). These pre-configured plots can be used in addition to your variant data plots to show the different signatures between your data and populations. We will discuss how to add these to your plots in your data in "Annotation and Visualization Parameters"> "General Options" section of the tutorial.

**Figure 8. Pre-configured Circos Plots**



AVIA's pre-generated Circos Plots

You are looking at AVIA's set of pre-configured Circos plots for Human based on population or CGI data. You may download these separately or add them to your own Circos Plots together with your data.

Download preLineRpt-preSimpleRpt-preAluRpt.png
Download hg19_CGI.snp135.1000k.png
Download preAluRpt-preLineRpt-preSimpleRpt.png
Download hg19.snp135-all.100k.png
Download preLineRpt-preSimpleRpt.png
Download USER_preAluRpt.png

## V) Workflows

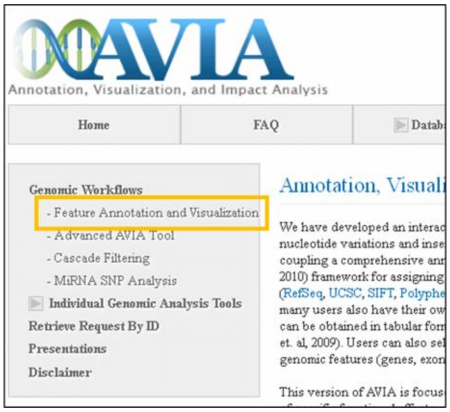
We will now describe the "Tools Navigation Panel" as highlighted in Figure 3c on the left side panel. Four workflows will be discussed: Feature Annotation and Visualization, Advanced AVIA Workflow, Cascade Filtering and miRNA SNP analysis. The Riken Liver cancer sample data may be used in 3 out of the 4 workflows to show how AVIA will report the variants differently. For the miRNA SNP analysis, there is a different usage example that you may use to run the workflow.

## A) Feature Annotation and Visualization

The annotation workflow is the driving component of AVIA. We will discuss major steps (data submission, parameters selection and data retrieval) of the AVIA workflows using the Riken data as an example.

### 1) Submission

**Figure 9. Navigating to AVIA Annotation Request**



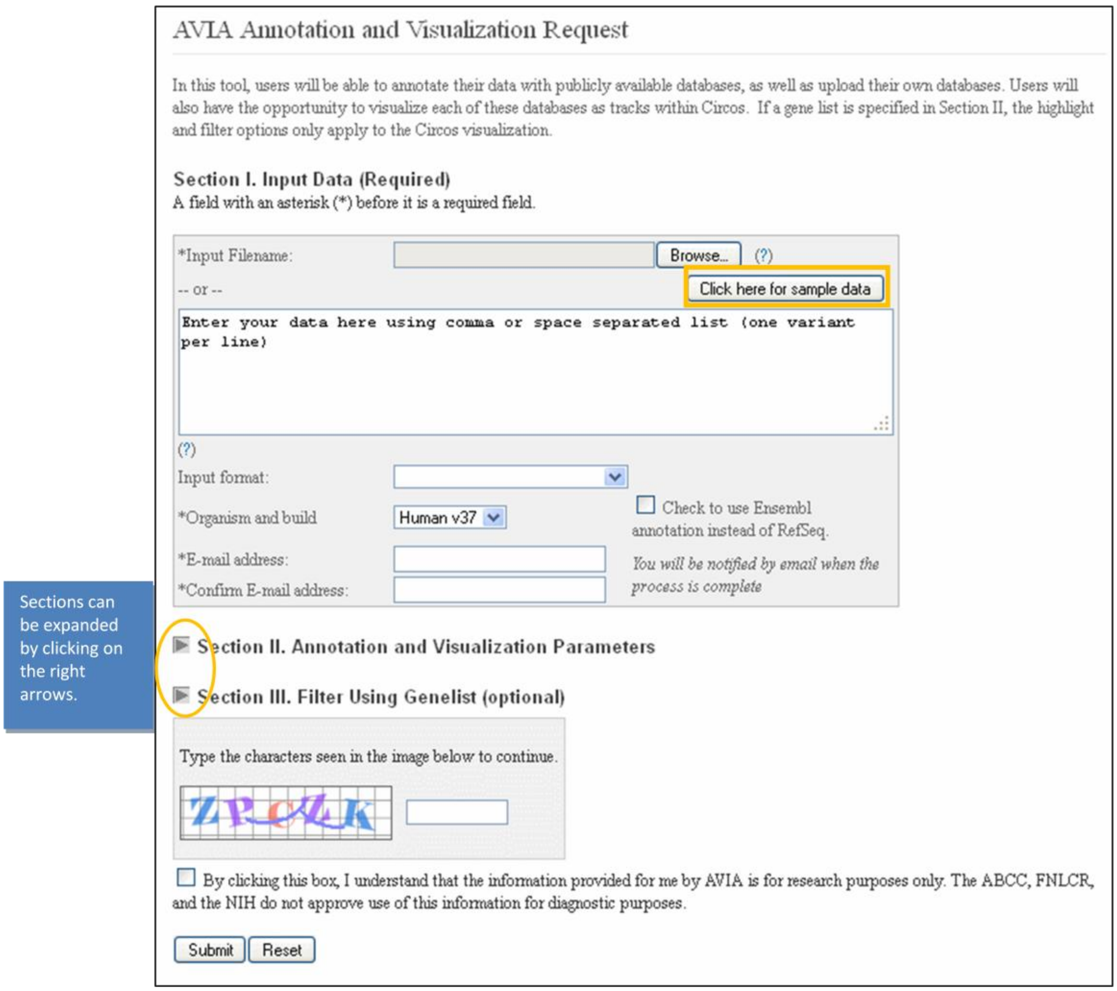
We have developed an interac
nucleotide variations and inse
coupling a comprehensive ann
2010) framework for assigning
(RefSeq, UCSC, SIFT, Polyphe
many users also have their ow
can be obtained in tabular forn
et. al, 2009). Users can also sel
genomic features (genes, exon

This version of AVIA is focus

To begin with a new annotation request, from the AVIA home page, you can navigate to the Genomic Workflows option at the Tools Navigation Panel, located on the left side of the webpage and click on the "Feature Annotation and Visualization" as highlighted in the orange box in Figure 9. Again, please enable javascript on your web browser; otherwise, your request will not submit or will result in an error message.

Clicking on the link will take you to the workflow page. There are three steps ("Sections") in this workflow that allow user to configure the annotation request. Section I defines the variant data, input format, origin of the data, the upload details, selection of the annotated gene features (RefSeq or Ensembl) and user's contact information. This Section is a required first step in any AVIA workflow. As described in the blue box of Figure 10 (next page), Sections II and III, clicking on the right arrows will expand the section. Section II identifies the annotation and visualization parameters of the request and Section III accommodates for gene-list focused annotation including user-supplied lists of genes and pathways. Below are some details of each of the three sections.

**Figure 10. Annotation and Visualization Request Page**



AVIA Annotation and Visualization Request

In this tool, users will be able to annotate their data with publicly available databases, as well as upload their own databases. Users will also have the opportunity to visualize each of these databases as tracks within Circos. If a gene list is specified in Section II, the highlight and filter options only apply to the Circos visualization.

**Section I. Input Data (Required)**
A field with an asterisk (*) before it is a required field.

*Input Filename: [Browse...] (?)
-- or -- [Click here for sample data]

```
Enter your data here using comma or space separated list (one variant
per line)
```

(?)
Input format:
*Organism and build: Human v37
□ Check to use Ensembl annotation instead of RefSeq.
*E-mail address:
*Confirm E-mail address:
*You will be notified by email when the process is complete*

Sections can be expanded by clicking on the right arrows.

▶ Section II. Annotation and Visualization Parameters

▶ Section III. Filter Using Genelist (optional)

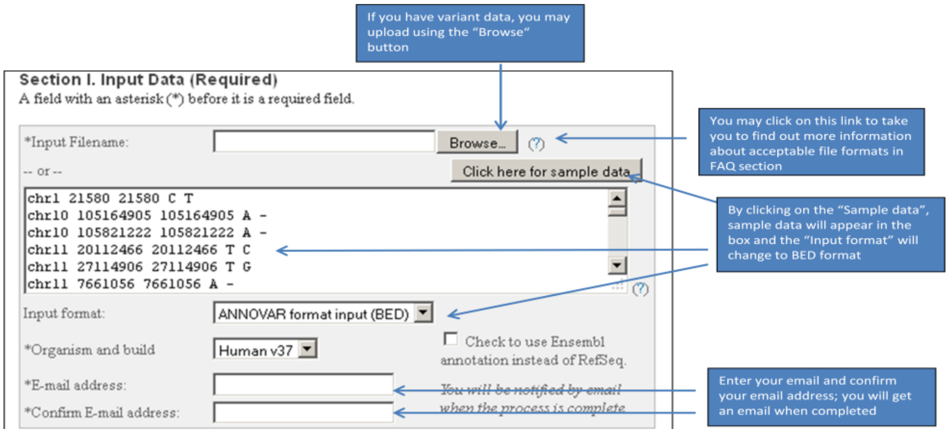
Type the characters seen in the image below to continue.

□ By clicking this box, I understand that the information provided for me by AVIA is for research purposes only. The ABCC, FNLCR, and the NIH do not approve use of this information for diagnostic purposes.

[Submit] [Reset]

Section I. The input data can be submitted through the text box form by “Copy-Paste” or the input file with the data can be uploaded following the procedure: (i) click on the “Browse” button and navigate to your file; (ii) select your file using the browser and click “OK”. Your path and file should appear in the text box next to the “Browse” button.

**Figure 11. Input Data Section**



If you have variant data, you may upload using the “Browse” button

**Section I. Input Data (Required)**
A field with an asterisk (*) before it is a required field.

*Input Filename: [Browse...] (?)
-- or -- [Click here for sample data]

```
chr1 21580 21580 C T
chr10 105164905 105164905 A -
chr10 105821222 105821222 A -
chr11 20112466 20112466 T C
chr11 27114906 27114906 T G
chr11 7661056 7661056 A -
```

(?)
Input format: ANNOVAR format input (BED)
*Organism and build: Human v37
□ Check to use Ensembl annotation instead of RefSeq.
*E-mail address:
*Confirm E-mail address:
~~You will be notified by email when the process is complete~~

You may click on this link to take you to find out more information about acceptable file formats in FAQ section

By clicking on the “Sample data”, sample data will appear in the box and the “Input format” will change to BED format

Enter your email and confirm your email address; you will get an email when completed

*TUTORIAL: If you do not have any data, you can still follow along by clicking on the "Click here for sample data" button, highlighted in orange, as in the Figure 11. Variation data will appear in the box or in the text box where you uploaded your file. For the sample data, the "Input format" will be changed to the "ANNOVAR format input (BED)".*

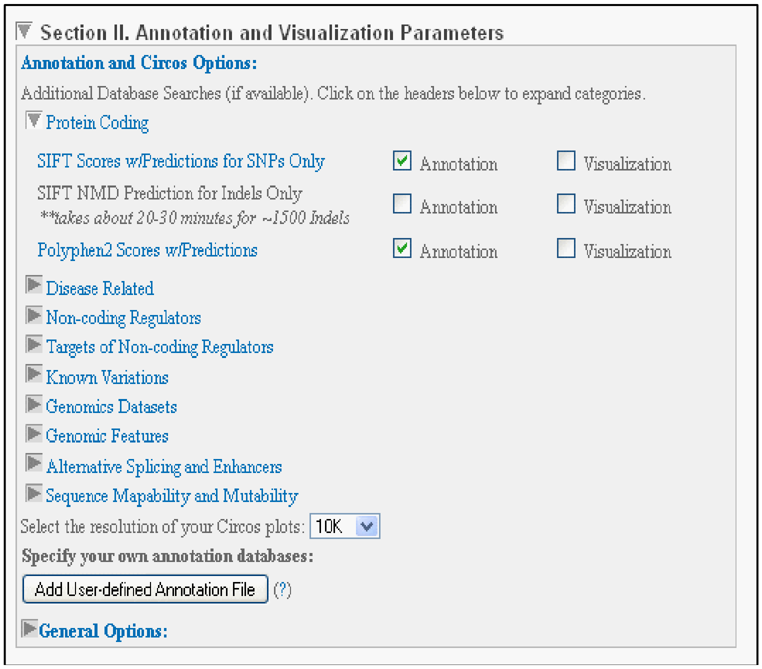
If you upload a VCF4 file or any other format, you must change the "Input format" or AVIA will error and will not run. Remember that all the formats will be converted to BED-link format. You may also click on the "(?)" next to the browse button to learn about input formats or visit Section III "File Formats" in this tutorial.

Next you should select an organism, genome assembly version applied to your data and basic annotation source you wish to use.

We currently support RefSeq Human genome references (v36 and v37) as the default annotation and Ensembl as an option (v63). If your input coordinates are an older build, please note that we will use UCSC's Liftover to convert the coordinates to the newest build. Your original coordinates will be retained in the final report. Complete Section I by filling in your email address.

*TUTORIAL: Please use "Human v37" as the Organism and Build. Do not check Ensembl box.*

**Figure 12. Annotation and Visualization Parameters**



▼ Section II. Annotation and Visualization Parameters

**Annotation and Circos Options:**

Additional Database Searches (if available). Click on the headers below to expand categories.

▼ Protein Coding

| | | |
|---|---|---|
| SIFT Scores w/Predictions for SNPs Only | ☑ Annotation | □ Visualization |
| SIFT NMD Prediction for Indels Only<br>***takes about 20-30 minutes for ~1500 Indels* | □ Annotation | □ Visualization |
| Polyphen2 Scores w/Predictions | ☑ Annotation | □ Visualization |

▶ Disease Related

▶ Non-coding Regulators

▶ Targets of Non-coding Regulators

▶ Known Variations

▶ Genomics Datasets

▶ Genomic Features

▶ Alternative Splicing and Enhancers

▶ Sequence Mapability and Mutability

Select the resolution of your Circos plots: 10K

**Specify your own annotation databases:**

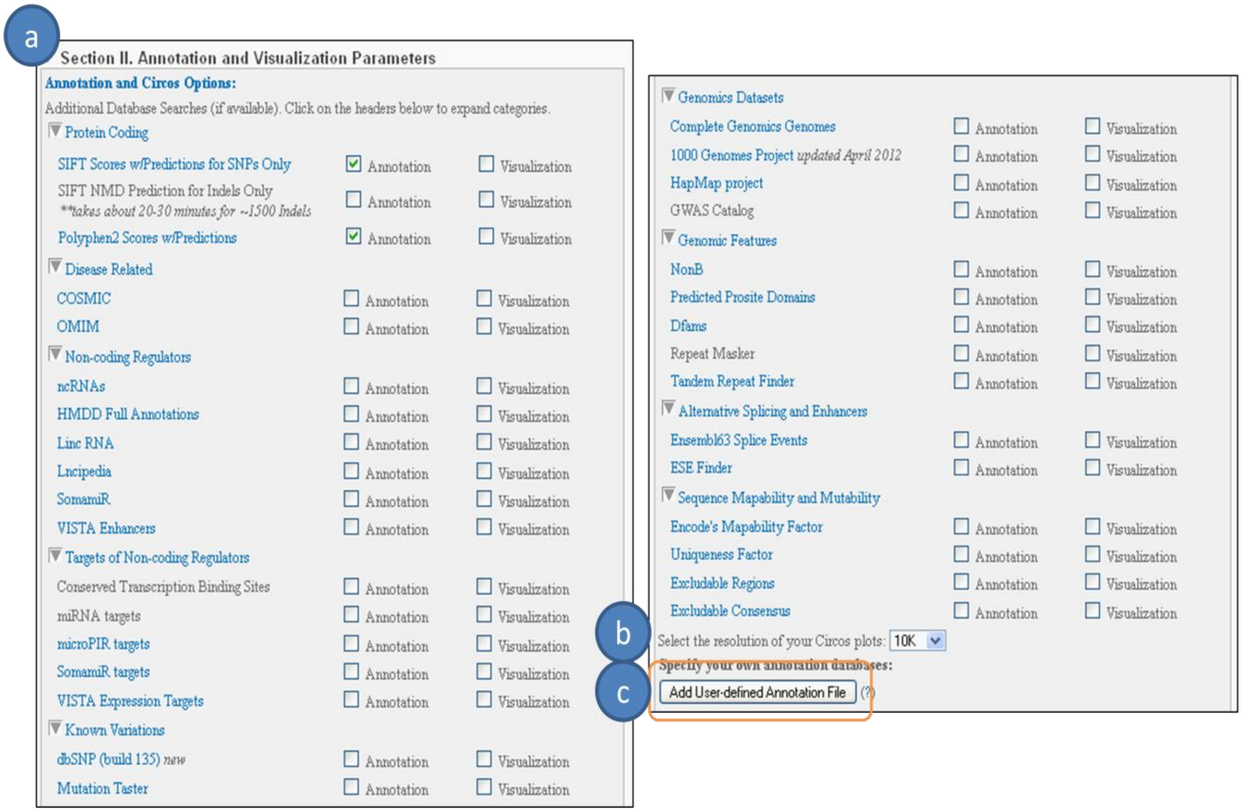
Add User-defined Annotation File (?)

▶ **General Options:**

Figure 12 shows a list of publicly available databases we put together for annotation of the variations data, described in more detail on our webpage in Figure 7. Figure 12 is the default screen of the Annotation and Visualization section when you first arrive on the page. By default, the "SIFT Scores w/Predictions" and "Polyphen2 Scores w/Prediction" Annotation buttons are checked. You may click on the ▶ icon to expand or ▼ collapse the headers.

Figure 13 is the expanded version of the page. The list of the databases in the screenshot may be different than the current version; however, it illustrates two options provided by AVIA for each database in the collection:  “Annotation” and “Visualization”.  Checking on “Annotation” box will result in additional annotation column reflecting the database content, if it exists. The “Visualization” feature can be run in addition to the “Annotation” and will lead to generation of the data track by Circos (http://www.Circos.ca/ ).

**Figure 13. Expanded Parameters for Annotation and Visualization**



Section II. Annotation and Visualization Parameters

**Annotation and Circos Options:**

Additional Database Searches (if available). Click on the headers below to expand categories.

▼ Protein Coding

SIFT Scores w/Predictions for SNPs Only ☑ Annotation ☐ Visualization

SIFT NMD Prediction for Indels Only **takes about 20-30 minutes for ~1500 Indels* ☐ Annotation ☐ Visualization

Polyphen2 Scores w/Predictions ☑ Annotation ☐ Visualization

▼ Disease Related

COSMIC ☐ Annotation ☐ Visualization

OMIM ☐ Annotation ☐ Visualization

▼ Non-coding Regulators

ncRNAs ☐ Annotation ☐ Visualization

HMDD Full Annotations ☐ Annotation ☐ Visualization

Linc RNA ☐ Annotation ☐ Visualization

Lncipedia ☐ Annotation ☐ Visualization

SomamiR ☐ Annotation ☐ Visualization

VISTA Enhancers ☐ Annotation ☐ Visualization

▼ Targets of Non-coding Regulators

Conserved Transcription Binding Sites ☐ Annotation ☐ Visualization

miRNA targets ☐ Annotation ☐ Visualization

microPIR targets ☐ Annotation ☐ Visualization

SomamiR targets ☐ Annotation ☐ Visualization

VISTA Expression Targets ☐ Annotation ☐ Visualization

▼ Known Variations

dbSNP (build 135) *new* ☐ Annotation ☐ Visualization

Mutation Taster ☐ Annotation ☐ Visualization

▼ Genomics Datasets

Complete Genomics Genomes ☐ Annotation ☐ Visualization

1000 Genomes Project *updated April 2012* ☐ Annotation ☐ Visualization

HapMap project ☐ Annotation ☐ Visualization

GWAS Catalog ☐ Annotation ☐ Visualization

▼ Genomic Features

NonB ☐ Annotation ☐ Visualization

Predicted Prosite Domains ☐ Annotation ☐ Visualization

Dfams ☐ Annotation ☐ Visualization

Repeat Masker ☐ Annotation ☐ Visualization

Tandem Repeat Finder ☐ Annotation ☐ Visualization

▼ Alternative Splicing and Enhancers

Ensembl63 Splice Events ☐ Annotation ☐ Visualization

ESE Finder ☐ Annotation ☐ Visualization

▼ Sequence Mapability and Mutability

Encode's Mapability Factor ☐ Annotation ☐ Visualization

Uniqueness Factor ☐ Annotation ☐ Visualization

Excludable Regions ☐ Annotation ☐ Visualization

Excludable Consensus ☐ Annotation ☐ Visualization

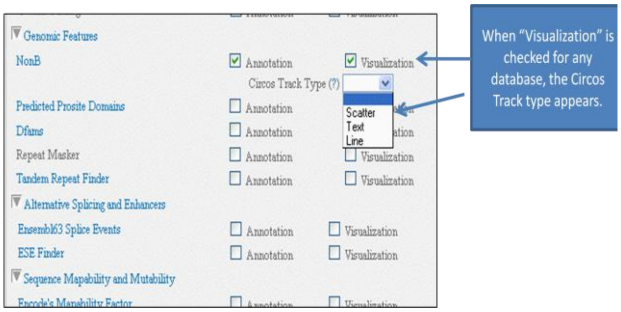
Select the resolution of your Circos plots: 10K

**Specify your own annotation databases:**

Add User-defined Annotation File (?)

Once you checked the “Visualization” option, you may then select the “Circos Track Type” (Figure 14) from the drop down box.  You will only have a limited track types due to the database.  For example, sparse data, such as Damaging SIFT or Polyphen Scores will be plotted using scatter or dots, while dbSNP can be plotted using lines or scatter. Circos tracks

**Figure 14. Circos Types**



▼ Genomic Features

NonB ☑ Annotation ☑ Visualization

Circos Track Type (?)

Predicted Prosite Domains ☐ Annotation

Dfams ☐ Annotation

Repeat Masker ☐ Annotation ☐ Visualization

Tandem Repeat Finder ☐ Annotation ☐ Visualization

▼ Alternative Splicing and Enhancers

Ensembl63 Splice Events ☐ Annotation ☐ Visualization

ESE Finder ☐ Annotation ☐ Visualization

▼ Sequence Mapability and Mutability

Encode's Mapability Factor ☐ Annotation ☐ Visualization

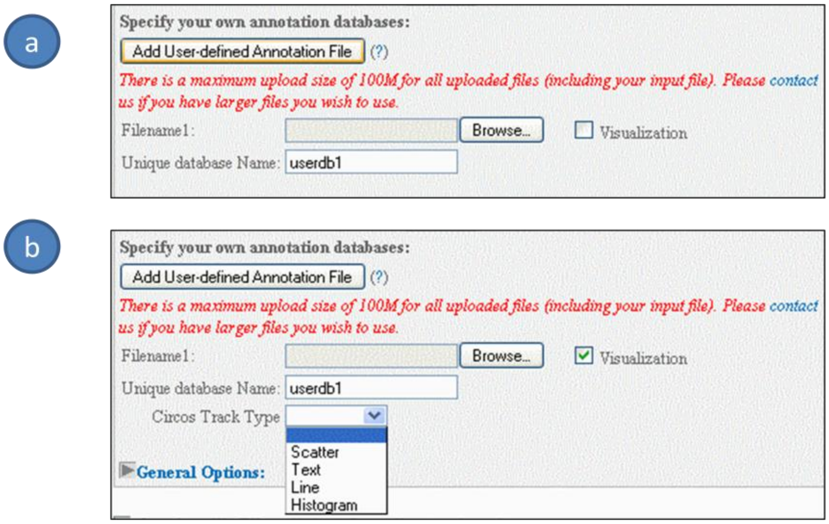
When “Visualization” is checked for any database, the Circos Track type appears.

corresponding to particular databases show percentage of genic variations per specified frame of the genome bin. The bin can be selected by the “Select the resolution of your plots” drop down menu found in the previous figure 13b.

If “Add User-defined Annotation File” was selected in Figure 13c, users may upload their own annotation data. By clicking the button (Figure 15a), options to upload and choose visualization options, and naming your database appear.  You may choose to upload up to 10 databases for annotation.  If you wish to plot a Circos plot using this database, click on the “Visualization” checkbox next to the “Browse” button, and “Circos Track Type” will appear as shown in Figure 15b.  The “Unique database Name” is a field which can be used to replace the header in the AVIA output (see “Headers” in Section V Output Interpretation).  This MUST be a unique name and cannot have any spaces or special characters (spaces, dashes, underscores, parentheses, brackets, ampersands, pound signs, punctuations, etc) in the name.

These details will appear each time the “Add User-defined Annotation File” button is pressed. Please note that you should first click to add the number of databases desired before filling in the details of each database.

**Figure 15.  User supplied Annotations**



a

Specify your own annotation databases:

*There is a maximum upload size of 100M for all uploaded files (including your input file). Please contact us if you have larger files you wish to use.*

Filename1:

□ Visualization

Unique database Name: userdb1

b

Specify your own annotation databases:

*There is a maximum upload size of 100M for all uploaded files (including your input file). Please contact us if you have larger files you wish to use.*

Filename1:

☑ Visualization

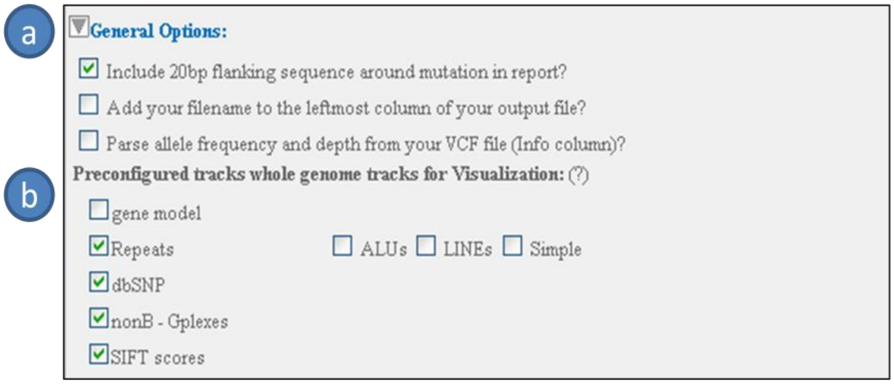
Unique database Name: userdb1

Circos Track Type

At the very bottom of Section II, there are several formatting options available as shown in Figure 16. These options pertain to general options, such as adding flanking sequence to the output file, adding original filename to the leftmost column of the output file and retrieving the allele frequency information from the VCF file. If the latter is selected and you submitted a valid VCF file, it will generate a column with the "AF=" and "DP=" tags in the info column of your VCF file. This option will be ignored if it is checked and you did not submit a VCF file.

Figure 16b refers to visualization of the preconfigured whole genome tracks on your results page. You may current view this data on http://aviadev.abcc.ncifcrf.gov/apps/site/successful_viz/?id=preconfigured_plots or by selecting "Databases"> "View preconfigured plots" on your User Navigation Panel at the top of any AVIA page. From that page however, the data cannot be integrated with user variants results. To add it to your results so they can be integrated, you must select them from this menu. This option is useful for your rearranging Circos Plots after your variants have been annotated and can be useful in comparing whole population datasets to your variant data.

**Figure 16. General Options for Annotation and Visualization and Cascade Filtering**



▼General Options:

☑ Include 20bp flanking sequence around mutation in report?

☐ Add your filename to the leftmost column of your output file?

☐ Parse allele frequency and depth from your VCF file (Info column)?

Preconfigured tracks whole genome tracks for Visualization: (?)

☐gene model

☑Repeats ☐ ALUs ☐ LINEs ☐ Simple

☑dbSNP

☑nonB - Gplexes

☑SIFT scores

The next section (Filter By Gene) allows users to focus (reduce) the annotation and visualization to specific set of the genic features. Figure 17 identifies 4 options of this section as denoted in the blue circles, which we describe below:

1) Filter vs. Highlight: This option refers to the Circos "text" track type visualization. You may wish to highlight specific genes in the variant set in such a way that your genes of interest (e.g in your gene list) will be colored red, and the rest of the mutations in genes will be colored black. Or you can choose to filter, which means only the genes in your desired gene list will be displayed, if there is a mutation in that gene.

2) Count by Gene vs Count by Variant: This option refers to how you wish to count the number of variants per bin. A bin will be defined in the "Annotation and Visualization" Section of the input parameters. If you wish to count each mutation regardless of whether it is in the same gene or not, choose, count by variants. In this option, you can have count for each bin mutation. If you wish to count mutations with genes, then choose count by

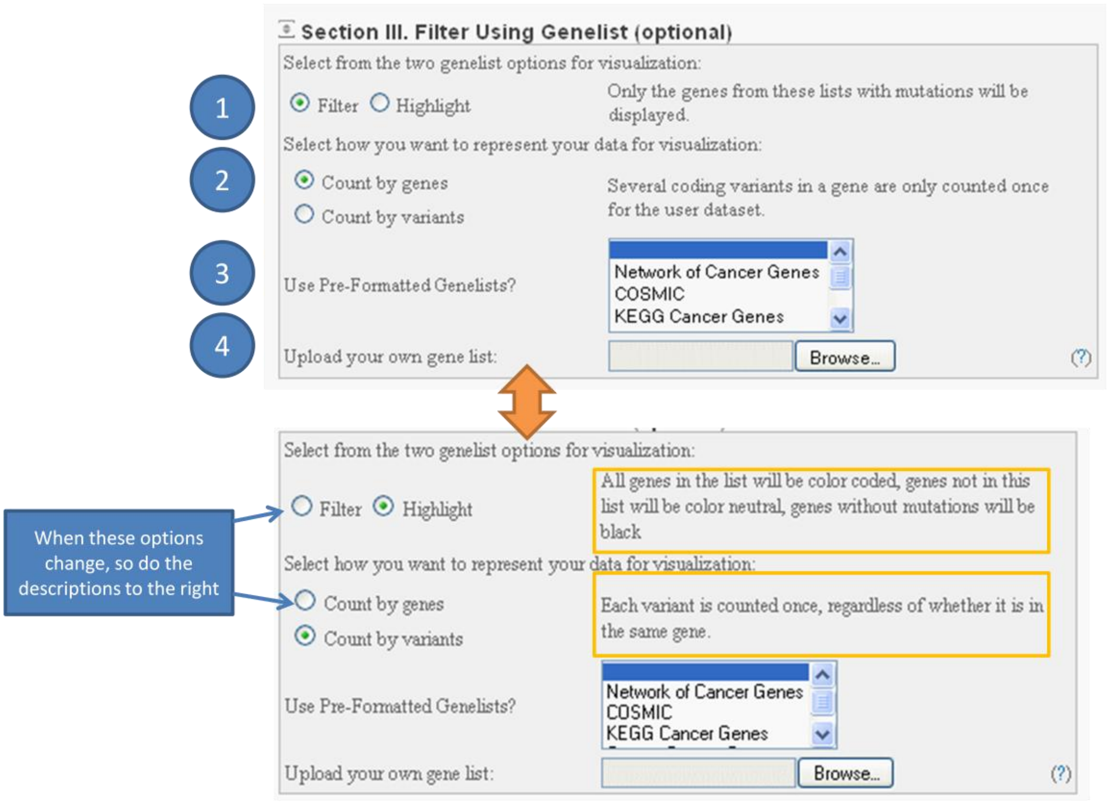
genes. This will mean that the gene mutation will count only once. This affects the frequency of the data being represented in the Circos plot.

*Note: As you change from "Filter" to "Highlight" or "Count by genes" to "Count by variants" or vice versa, the description to the right will change as shown in the second box with "Highlight" and "Count By Variants" selected.*

3) Use Pre-formatted Gene Lists: This option allows you to pick a gene list based on certain pathways. You may pick one or more gene lists in the option window. The gene lists are named for the source of the gene list. Your variants in these genes will be shown in the final reports. If you wish to select multiple gene lists, hold down the "CTRL" key on your keyboard and click on each of the gene lists preferred.

4) Upload your own gene list: This option allows you to upload your own gene list, one gene per row. If you have synonyms for those genes, it is also good to include those as well to ensure that all genes that you wish to be displayed are included in your final dataset. Like the previous option, this will filter for genes in your gene list and only display those requested.

**Figure 17. Gene List Filters Options**


1
2
3
4
When these options change, so do the descriptions to the right

Section III. Filter Using Genelist (optional)

Select from the two genelist options for visualization:

● Filter ○ Highlight — Only the genes from these lists with mutations will be displayed.

Select how you want to represent your data for visualization:

● Count by genes
○ Count by variants

Several coding variants in a gene are only counted once for the user dataset.

Use Pre-Formatted Genelists? Network of Cancer Genes / COSMIC / KEGG Cancer Genes

Upload your own gene list: Browse... (?)

Select from the two genelist options for visualization:

○ Filter ● Highlight — All genes in the list will be color coded, genes not in this list will be color neutral, genes without mutations will be black

Select how you want to represent your data for visualization:

○ Count by genes
● Count by variants

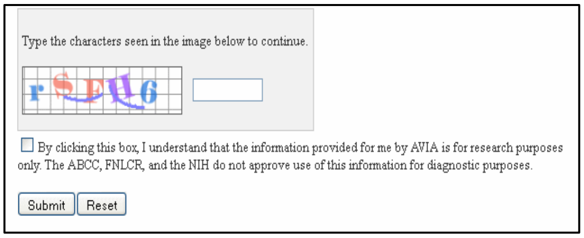
Each variant is counted once, regardless of whether it is in the same gene.

Use Pre-Formatted Genelists? Network of Cancer Genes / COSMIC / KEGG Cancer Genes

Upload your own gene list: Browse... (?)

To complete your AVIA submission request, you must enter the “Captcha” in the last grey box and read and check the disclaimer acknowledgement before clicking on the “Submit” button, shown in Figure 18.

**Figure 18. Captcha and Disclaimer**



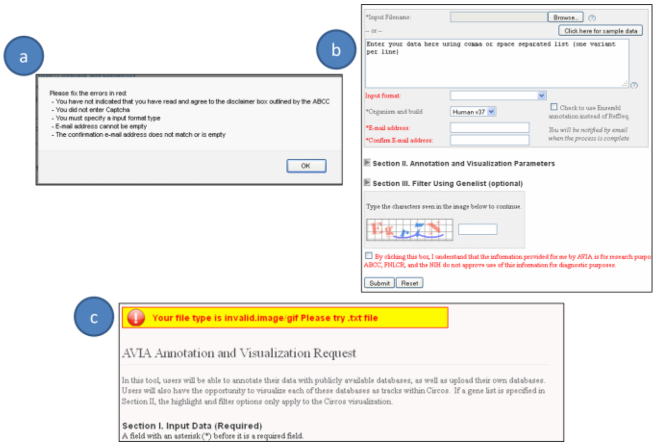
Type the characters seen in the image below to continue.

□ By clicking this box, I understand that the information provided for me by AVIA is for research purposes only. The ABCC, FNLCR, and the NIH do not approve use of this information for diagnostic purposes.

If there are any errors with your submission, a pop-up box will appear with the list of errors detected and the errant Sections on the webpage will be highlighted in red font (Figure 19a,b).

If there is a problem with your file upload(s), AVIA will provide an error message in a yellow banner as shown in Figure 19c. There is an upload limit of 200MB per file. You will get this error if your input file is too big or not the correct file format, as determined by the input file’s suffix. In the example in Figure 19c, an invalid image file was uploaded.

**Figure 19. Submission Errors**



Please fix the errors in red:
- You have not indicated that you have read and agree to the disclaimer box outlined by the ABCC
- You did not enter Captcha
- You must specify a input format type
- E-mail address cannot be empty
- The confirmation e-mail address does not match or is empty

*Input Filename:

-- or --

Enter your data here using comma or space separated list (one variant per line)

Input format:

*Organism and build: Human v37

□ Check to use Ensembl annotation instead of RefSeq

*E-mail address:

*Confirm E-mail address:

You will be notified by email when the process is complete

Section II. Annotation and Visualization Parameters

Section III. Filter Using Genelist (optional)

Type the characters seen in the image below to continue.

□ By clicking this box, I understand that the information provided for me by AVIA is for research purpo
ABCC, FNLCR, and the NIH do not approve use of this information for diagnostic purposes.

Your file type is invalid.image/gif Please try .txt file

AVIA Annotation and Visualization Request

In this tool, users will be able to annotate their data with publicly available databases, as well as upload their own databases. Users will also have the opportunity to visualize each of these databases as tracks within Circos. If a gene list is specified in Section II, the highlight and filter options only apply to the Circos visualization.

Section I. Input Data (Required)

A field with an asterisk (*) before it is a required field.

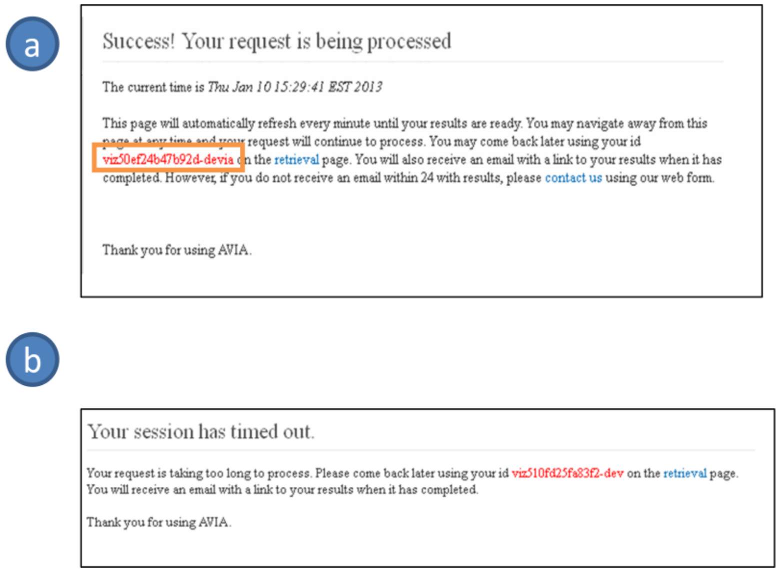
## 2) Data Retrieval

Results retrieval is a multi-step process which begins with processing page and then results in the “Results” page. Here we discuss what happens after you click the “Submit” button and how to retrieve data. The results pages will be discussed in Section VI Output Interpretation.

### a) Processing

If you have successful submitted your request, you will be redirected to the status page as shown in Figure 20a. This page is updated every minute until your request is complete, up to 5 minutes. If your request completes within the allotted time, you will be automatically redirected to your results page. However, if your request takes longer than 5 minutes, you will be directed on how to proceed as shown in Figure 20b. If you choose to navigate away from this page, please take note of the unique identifier given to you , which is highlighted in orange in red font in Figure 20a. The time to process your request will depend on the size of the request and number of requests submitted to our servers.

**Figure 20. Processing Page**



Success! Your request is being processed

The current time is *Thu Jan 10 15:29:41 EST 2013*

This page will automatically refresh every minute until your results are ready. You may navigate away from this page at any time and your request will continue to process. You may come back later using your id viz50ef24b47b92d-devia on the retrieval page. You will also receive an email with a link to your results when it has completed. However, if you do not receive an email within 24 with results, please contact us using our web form.

Thank you for using AVIA.

Your session has timed out.

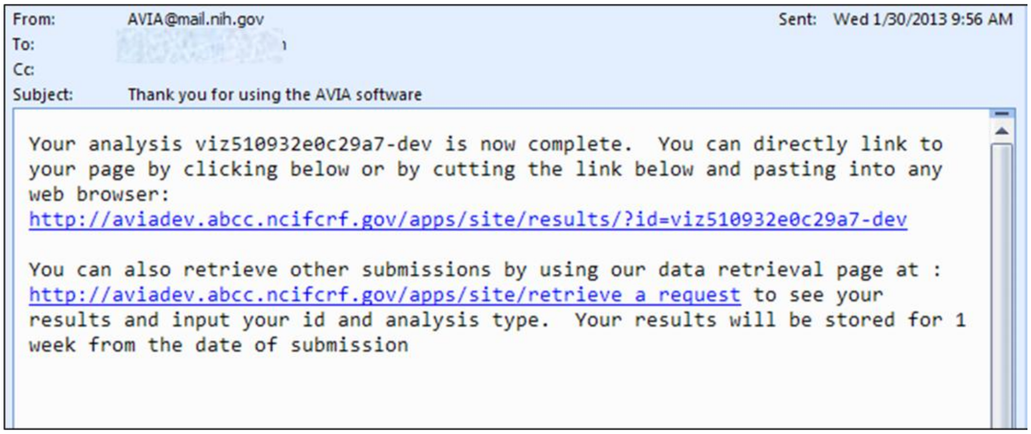
Your request is taking too long to process. Please come back later using your id viz510fd25fa83f2-dev on the retrieval page. You will receive an email with a link to your results when it has completed.

Thank you for using AVIA.

## b) Email notification

You will also be notified by email at the email address provided upon completion of the request with a direct link to your results (Figure 21). Before contacting us, please check your spam folder if you do not receive a notification within 6 hours. You will have a week to retrieve your data.

**Figure 21. Email notification**



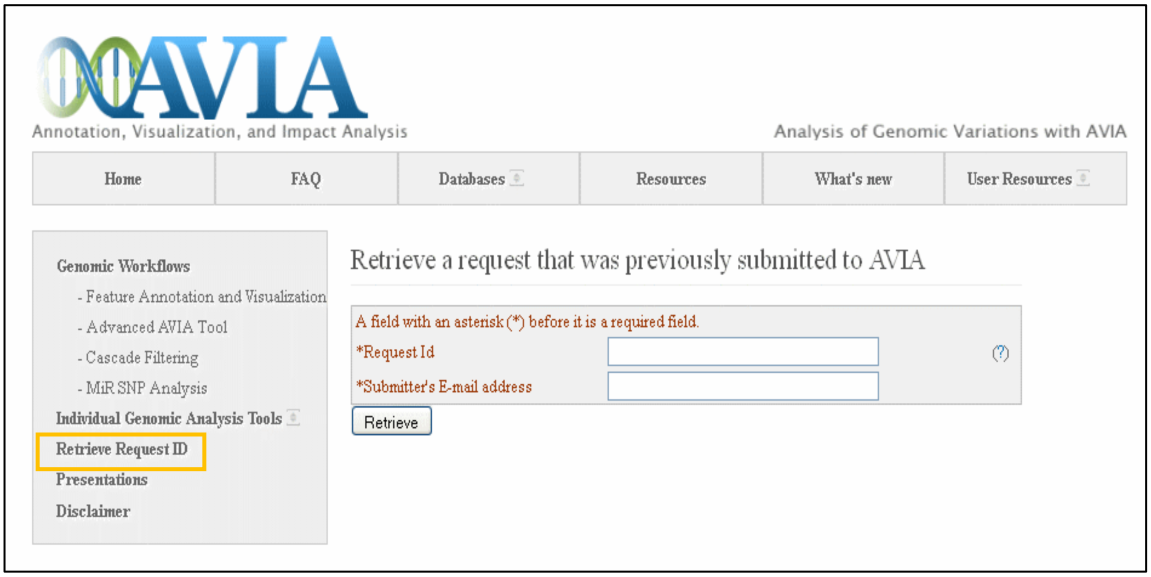
From: AVIA@mail.nih.gov
Sent: Wed 1/30/2013 9:56 AM
To: [redacted]
Cc:
Subject: Thank you for using the AVIA software

```
Your analysis viz510932e0c29a7-dev is now complete.  You can directly link to
your page by clicking below or by cutting the link below and pasting into any
web browser:
http://aviadev.abcc.ncifcrf.gov/apps/site/results/?id=viz510932e0c29a7-dev

You can also retrieve other submissions by using our data retrieval page at :
http://aviadev.abcc.ncifcrf.gov/apps/site/retrieve_a_request to see your
results and input your id and analysis type.  Your results will be stored for 1
week from the date of submission
```

### c) Request Retrieval Page

To retrieve a request after it has timed out or if you have navigated away, click on the "Retrieve Request" from the Tools menu on the left side. If you do not have the left menu, click on the AVIA logo and it will take you to the default page. Once on the page shown in Figure 22, you can enter the unique identifier from above and the email used when you submitted your request. If completed, it will take you to your data. If not, please return later. If you have not received an email from us within 6 hours, please contact us using our web form under "User Resources"> "Contact Us".

**Figure 22. Retrieve Request Page**



Retrieve a request that was previously submitted to AVIA

A field with an asterisk (*) before it is a required field.

*Request Id

*Submitter's E-mail address

We will discuss viewing results and interpreting the files in Section VI Output Interpretation of this tutorial.

## B) Cascade Filtering (Data Reduction) Method

Previous versions of ANNOVAR's auto_annotate.pl focused on a coding-centric strategy which could potentially eliminate key variants involved in disease. Originally the idea was to be able to reduce the amount of variants to novel mutations that could be further studied. However, by eliminating key variants, like those found in dbSNP, you may also be eliminating disease causing mutations. ANNOVAR's auto_annotate.pl script used a pre-selected set of databases which could not be altered in the pipeline. Using our cascade filtering data reduction method, the user controls with which databases to filter, which data to keep, and which cutoffs (if any) to use. From the remaining set, annotation and visualization is still performed. You may choose to filter a dataset using a SIFT above a cutoff value of 0.05, and then still choose to annotate SIFT to see what the cutoff value was.

To use the cascade filtering data reduction method, click on the "Cascade Filtering" on the AVIA Tools Panel located on the left side of any AVIA page, indicated by the orange arrow and box in Figure 23.

This will take you to a page that is very similar to the Feature Annotation and Visualization page (Figure 24); however notice that "Section II" is now Cascade filtering Parameters. The section for Input Data is still required and "Annotation and Visualization Parameters Section" is also available for use.

Figure 23. Navigating to Cascade Filtering Workflow

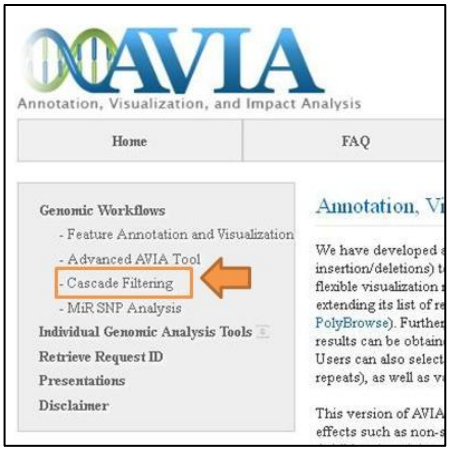
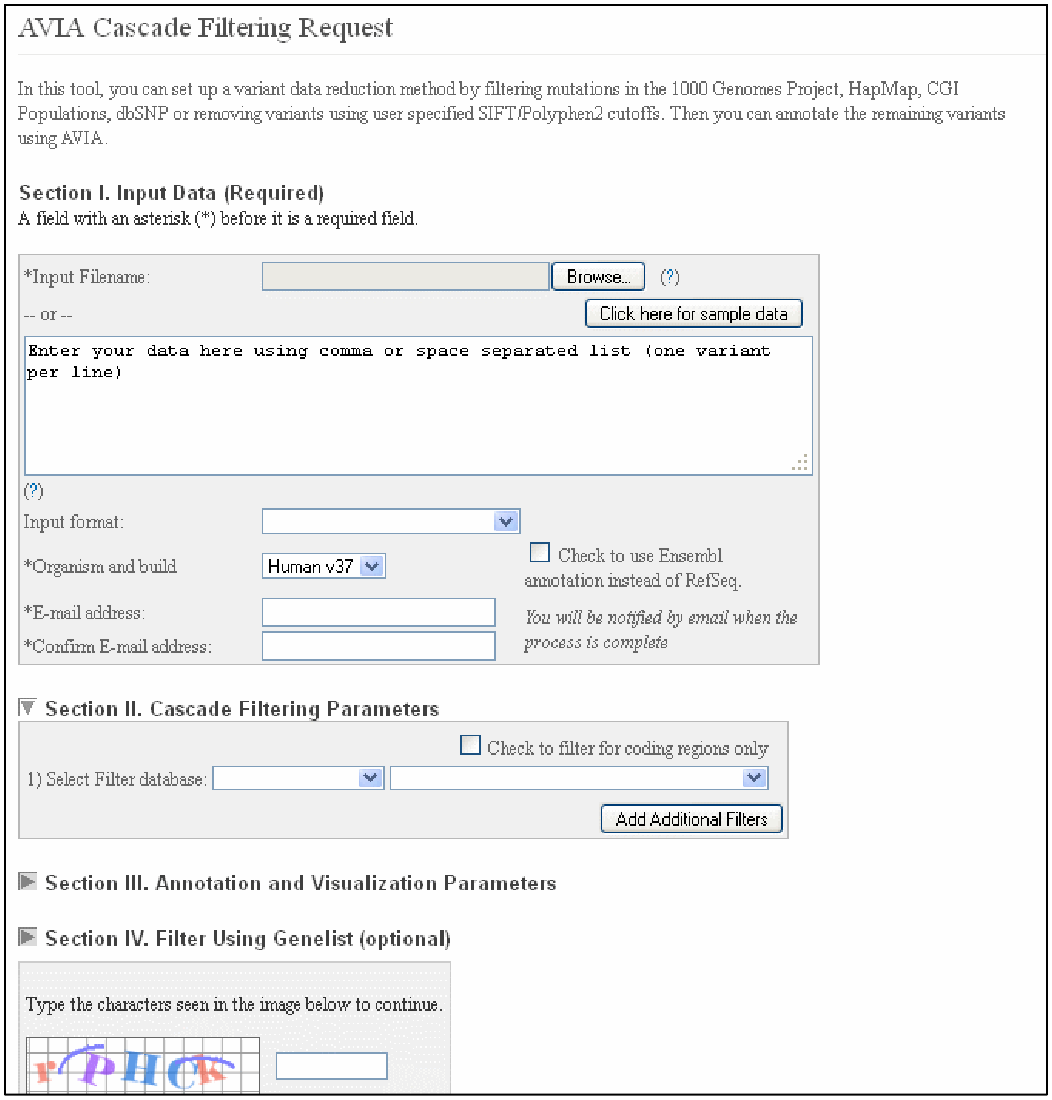


**Figure 24. Cascade Filtering Workflow**



AVIA Cascade Filtering Request

In this tool, you can set up a variant data reduction method by filtering mutations in the 1000 Genomes Project, HapMap, CGI Populations, dbSNP or removing variants using user specified SIFT/Polyphen2 cutoffs. Then you can annotate the remaining variants using AVIA.

Section I. Input Data (Required)

A field with an asterisk (*) before it is a required field.

*Input Filename: Browse... (?)

-- or -- Click here for sample data

```
Enter your data here using comma or space separated list (one variant
per line)
```

(?)

Input format:

*Organism and build Human v37

□ Check to use Ensembl annotation instead of RefSeq.

*E-mail address:

*Confirm E-mail address:

*You will be notified by email when the process is complete*

Section II. Cascade Filtering Parameters

□ Check to filter for coding regions only

1) Select Filter database:

Add Additional Filters

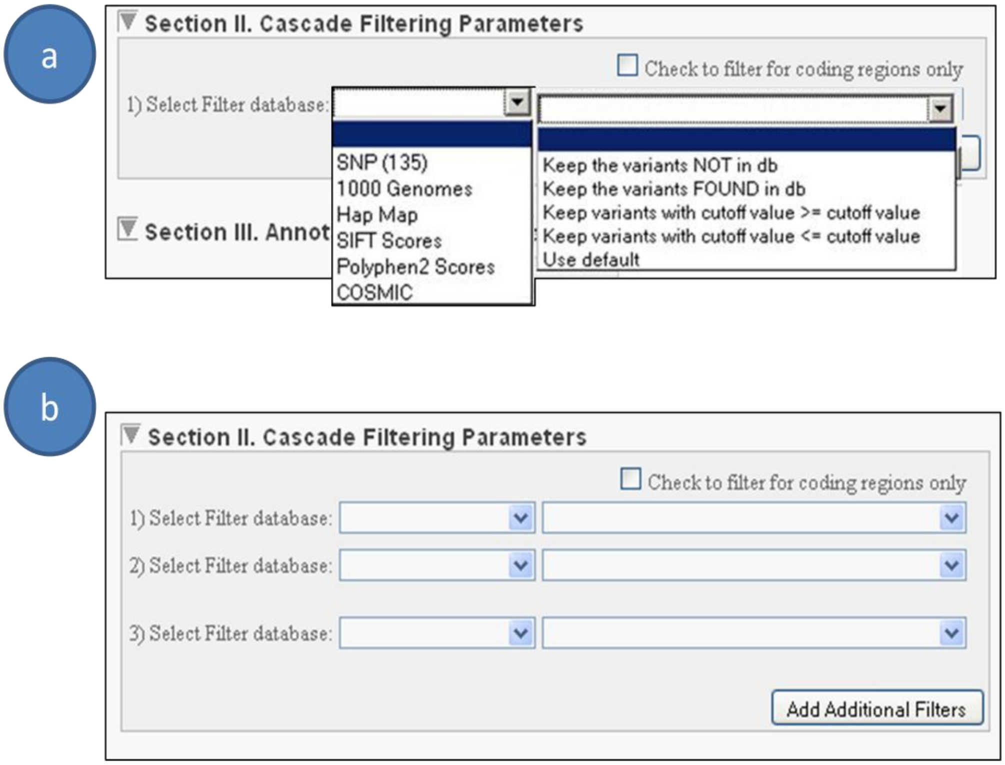
Section III. Annotation and Visualization Parameters

Section IV. Filter Using Genelist (optional)

Type the characters seen in the image below to continue.

In Figure 25, (a) you can choose which databases to filter with and set the cutoffs and which variants to keep and (b) you can add multiple filters up to 6 from the drop down menu list.  You may then continue to select Annotations and Visualization Parameters and Gene Filter Options in Sections III and IV, respectively as shown in Figure 24 (collapsed).

**Figure 25.  Cascade Filtering Options**



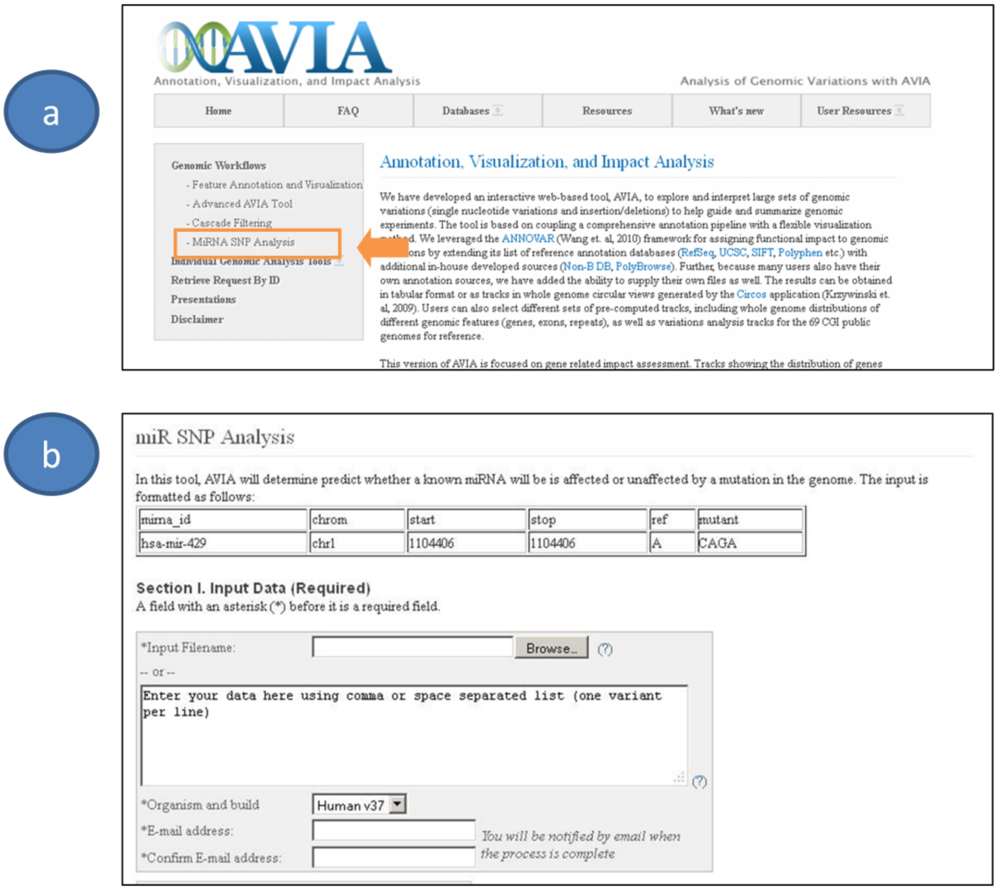
a
Section II. Cascade Filtering Parameters
□ Check to filter for coding regions only
1) Select Filter database:
SNP (135)
1000 Genomes
Hap Map
SIFT Scores
Polyphen2 Scores
COSMIC
Keep the variants NOT in db
Keep the variants FOUND in db
Keep variants with cutoff value >= cutoff value
Keep variants with cutoff value <= cutoff value
Use default
Section III. Annot
b
Section II. Cascade Filtering Parameters
□ Check to filter for coding regions only
1) Select Filter database:
2) Select Filter database:
3) Select Filter database:
Add Additional Filters

## C) miRNA SNP Analysis

This pipeline is slightly different from the previous ones in that the impact analysis of miR SNPs are calculated by each individual mutation to determine whether or not the change in nucleotide will affect the seed region of the miRNA.  miRNAs are very important in negative regulation in the cell and changes in the ability to bind to targets may adversely affect the interactions within the cell.

To begin, click on the “miR SNP Analysis” link on the left navigation panel as shown in Figure 25a. The input to this page requires *a priori* knowledge of the region that you are requesting impact analysis.  In the Figure 25b example at the top of the page, the mirna id should be added to the beginning of each line, followed by the chromosome, start and stop positions, reference allele and variant allele.  You can choose to either upload a tab delimited file or input the values in the “Cut-and-paste” box.  After selecting the Organism and Build and entering Email addresses to the form, you may submit.

**Figure 26. miRNA SNP Analysis**



Annotation, Visualization, and Impact Analysis

We have developed an interactive web-based tool, AVIA, to explore and interpret large sets of genomic variations (single nucleotide variations and insertion/deletions) to help guide and summarize genomic experiments. The tool is based on coupling a comprehensive annotation pipeline with a flexible visualization method. We leveraged the ANNOVAR (Wang et. al, 2010) framework for assigning functional impact to genomic variations by extending its list of reference annotation databases (RefSeq, UCSC, SIFT, Polyphen etc.) with additional in-house developed sources (Non-B DB, PolyBrowse). Further, because many users also have their own annotation sources, we have added the ability to supply their own files as well. The results can be obtained in tabular format or as tracks in whole genome circular views generated by the Circos application (Krzywinski et. al, 2009). Users can also select different sets of pre-computed tracks, including whole genome distributions of different genomic features (genes, exons, repeats), as well as variations analysis tracks for the 69 CGI public genomes for reference.

This version of AVIA is focused on gene related impact assessment. Tracks showing the distribution of genes

miR SNP Analysis

In this tool, AVIA will determine predict whether a known miRNA will be is affected or unaffected by a mutation in the genome. The input is formatted as follows:

| mirna_id | chrom | start | stop | ref | mutant |
|---|---|---|---|---|---|
| hsa-mir-429 | chr1 | 1104406 | 1104406 | A | CAGA |

Section I. Input Data (Required)

A field with an asterisk (*) before it is a required field.

*Input Filename:

-- or --

```
Enter your data here using comma or space separated list (one variant
per line)
```

*Organism and build: Human v37

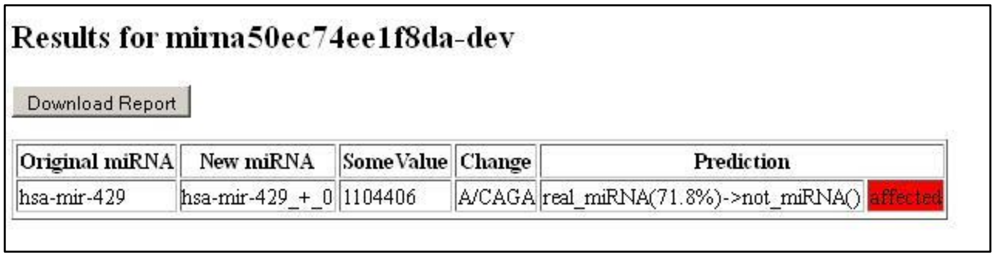
*E-mail address:

*Confirm E-mail address:

*You will be notified by email when the process is complete*

As seen in Figure 27, the output of this pipeline is a simple affected/not affected result. From here, you can see which miRNAs may be affected by the mutation.

**Figure 27. miRNA SNP Analysis Results Page**



Results for mirna50ec74ee1f8da-dev

| Original miRNA | New miRNA | Some Value | Change | Prediction | |
|---|---|---|---|---|---|
| hsa-mir-429 | hsa-mir-429_+_0 | 1104406 | A/CAGA | real_miRNA(71.8%)->not_miRNA() | affected |

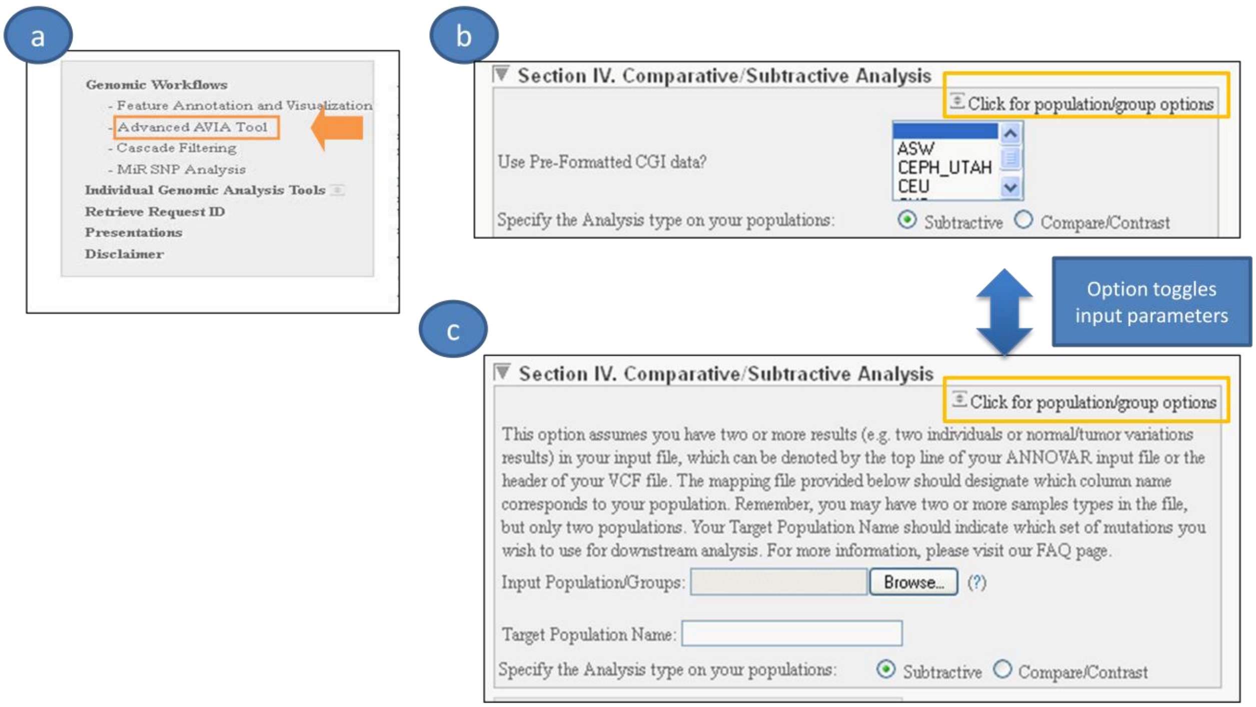
## D) Advanced AVIA Workflow

This workflow has an additional component called Subtractive analysis, which is a type of filtering using related variant data. The databases we use for annotation describe a mutation in the whole genome context, e.g. dbSNP are known mutation sets across multiple populations, annotation of miRNA targets and SIFT scores to predict damaging mutations, etc are calculated based on the reference genomes. Subtractive analysis is more focused towards personalized impact assessment. We describe here how this workflow can be used to derive novel mutations within one individual or sub-population. For example, i) an individual whose variant data were sequenced from two sources, e.g tumor vs normal or ii) disease from parental lineage, e.g. child vs parental can be found. This is another way to reduce the complexity of the data, but in a more meaningful way. This workflow can also be applied to populations. If the variant data was from an individual who was sequenced and was of a different ethnicity than the NCBI reference genome, we can subtract that individual's race-specific mutations by using the allele frequencies in the Complete Genomics Population data set.

For this tutorial, we refer to "target" population as the population in which the mutations will be kept. In the examples used above, these will be the tumor sequenced mutations, child mutations or the individual's mutations. The "normal" datasets will be the blood mutations, parental and healthy ethnic populations (CEU, JPT, MXL, etc).

To begin, click on the "Advanced AVIA Workflow" on the AVIA tools panel as shown Figure 28a and fill in Sections I-III as described above. For Section IV "Comparative /Subtractive Analysis", you have the option designate the ethnicity of your normal population based on CGI populations available. You also have the option to toggle between using CGI populations as the normal population, or to

**Figure 28. Advanced AVIA Workflow**


a
Genomic Workflows
- Feature Annotation and Visualization
- Advanced AVIA Tool
- Cascade Filtering
- MiR SNP Analysis
Individual Genomic Analysis Tools
Retrieve Request ID
Presentations
Disclaimer
b
Section IV. Comparative/Subtractive Analysis
Click for population/group options
ASW
CEPH_UTAH
CEU
c
Section IV. Comparative/Subtractive Analysis
Click for population/group options
Browse...
(?)

Use Pre-Formatted CGI data?

Specify the Analysis type on your populations: ● Subtractive ○ Compare/Contrast

Option toggles input parameters

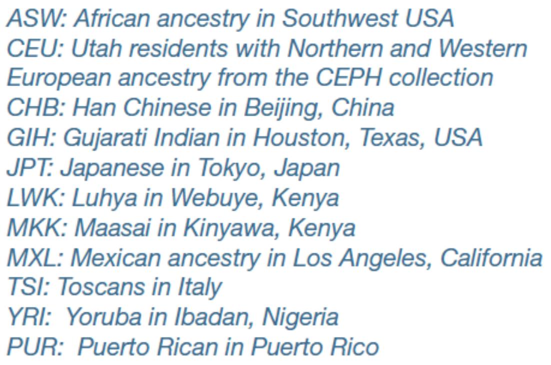
This option assumes you have two or more results (e.g. two individuals or normal/tumor variations results) in your input file, which can be denoted by the top line of your ANNOVAR input file or the header of your VCF file. The mapping file provided below should designate which column name corresponds to your population. Remember, you may have two or more samples types in the file, but only two populations. Your Target Population Name should indicate which set of mutations you wish to use for downstream analysis. For more information, please visit our FAQ page.

Input Population/Groups:

Target Population Name:

Specify the Analysis type on your populations: ● Subtractive ○ Compare/Contrast upload your own data to use as normal as shown in Figure 28 b and c. The ethnicity categories were taken from Complete Genomics manual (Figure 29)

**Figure 29. Ethnicity Data from Complete Genomics**



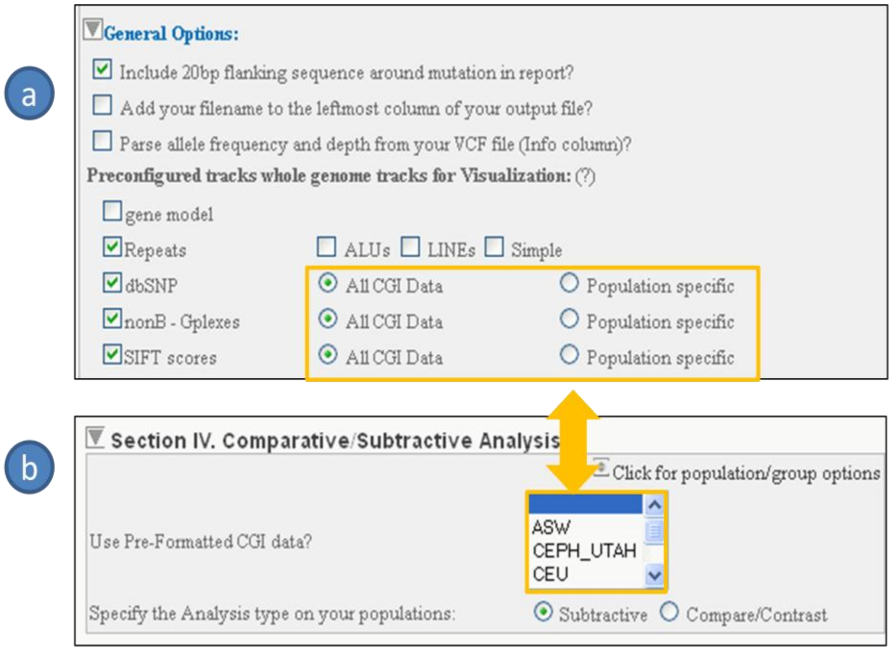
*ASW: African ancestry in Southwest USA*
*CEU: Utah residents with Northern and Western European ancestry from the CEPH collection*
*CHB: Han Chinese in Beijing, China*
*GIH: Gujarati Indian in Houston, Texas, USA*
*JPT: Japanese in Tokyo, Japan*
*LWK: Luhya in Webuye, Kenya*
*MKK: Maasai in Kinyawa, Kenya*
*MXL: Mexican ancestry in Los Angeles, California*
*TSI: Toscans in Italy*
*YRI: Yoruba in Ibadan, Nigeria*
*PUR: Puerto Rican in Puerto Rico*

If you choose to use your own normal population, it must be in the input VCF4 format, where the two populations are in the final 2 columns. Variants found in both your input file and your normal population will be removed from final variants set before Annotation and Visualization Parameters and Gene List Filters are applied, Sections II and III on the web page, respectively.

*TUTORIAL: For the sample data, fill out Section I with the sample data, and in Section III, you can use the CGI data and pick "JPT" in the selected panel on the left. The sample data was taken from the Riken Liver Cancer data from Japanese individuals. Continue to choose the annotation that you desire and click submit. The results will be exactly the same as the results for the Annotation pipeline save for the reduction of the population mutations found in both the normal and the sample data. (Figure 28c)*

There is one difference in the General Options section for the Advanced Workflow and the other workflows. As you can see from Figure 30a, now if you select dbSNP, Repeats, nonB or SIFT scores from the preconfigured visualization tracks, you can also choose to select which population for Circos visualization and comparision. This is either all CGI data combined, or population specific

**Figure 30. General Options for Advanced AVIA Workflow**



General Options:

☑ Include 20bp flanking sequence around mutation in report?
□ Add your filename to the leftmost column of your output file?
□ Parse allele frequency and depth from your VCF file (Info column)?

Preconfigured tracks whole genome tracks for Visualization: (?)

□ gene model
☑ Repeats □ ALUs □ LINEs □ Simple
☑ dbSNP ● All CGI Data ○ Population specific
☑ nonB - Gplexes ● All CGI Data ○ Population specific
☑ SIFT scores ● All CGI Data ○ Population specific

Section IV. Comparative/Subtractive Analysis

Click for population/group options

Use Pre-Formatted CGI data? ASW, CEPH_UTAH, CEU

Specify the Analysis type on your populations: ● Subtractive ○ Compare/Contrast ethnicity. If “Population specific” is selected, AVIA will automatically use the same ethnicity selected in Section IV of the website (Figure 30b). If none are selected, AVIA will default to “All CGI Data”

# VI) Output Interpretation

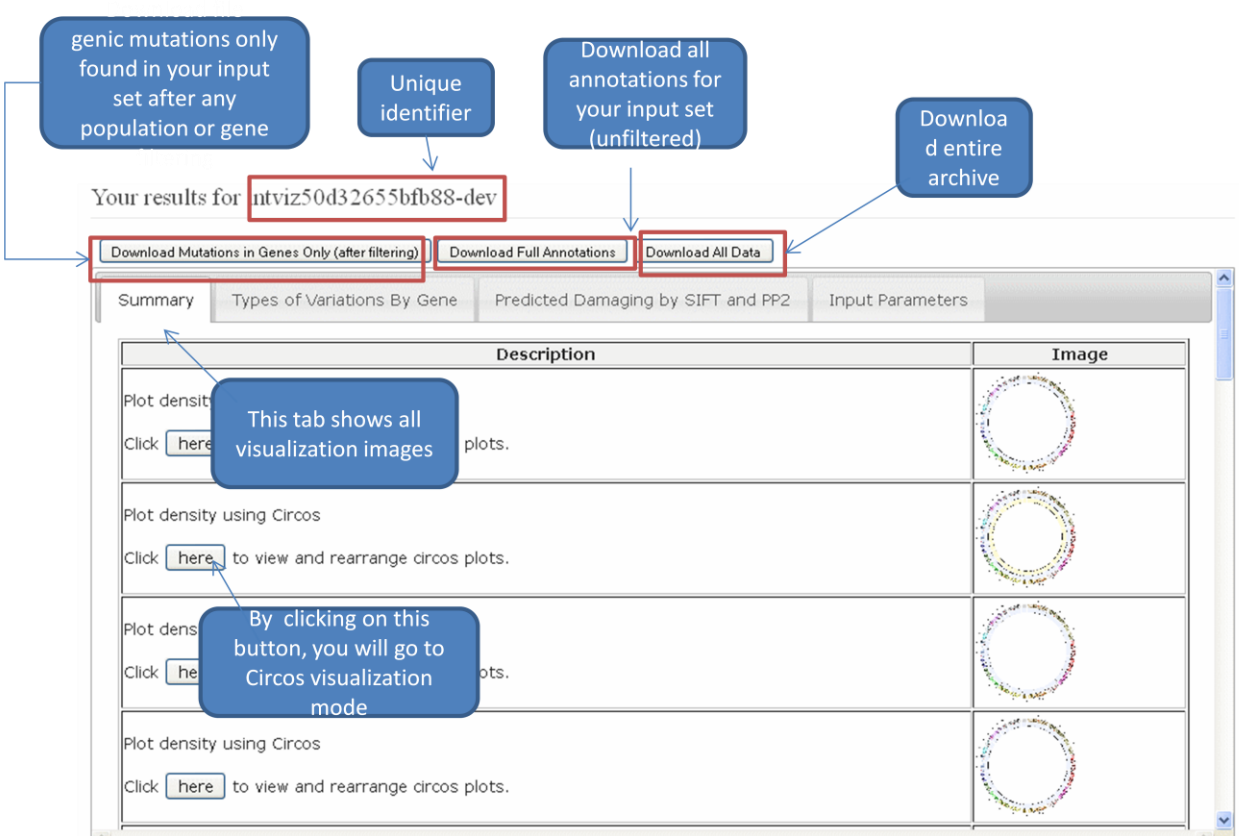
There are two ways to view the results, on the web or download to your local computer. In this section, we will describe the results as they are shown on the web and additional features that can be performed on the web. Your files will only remain on our server for 1 week after submission. At any time, you can download and save your data. In Sub section A, we will show how the files appear on the web, and in Sub section B, we will describe the final outputs.

## A) Viewing Results on the Web

### 1) Results for Annotation and Visualization, Cascade Filtering, and Subtractive Analysis Workflows

For all workflows except for the “miR SNP Analysis” workflow, the results pages are very similar. Figure 31 is a diagram of the results landing page.

**Figure 31. Results Landing Page**



Download file genic mutations only found in your input set after any population or gene filtering

Unique identifier

Download all annotations for your input set (unfiltered)

Downloa d entire archive

This tab shows all visualization images

By clicking on this button, you will go to Circos visualization mode

Your results for ntviz50d32655bfb88-dev

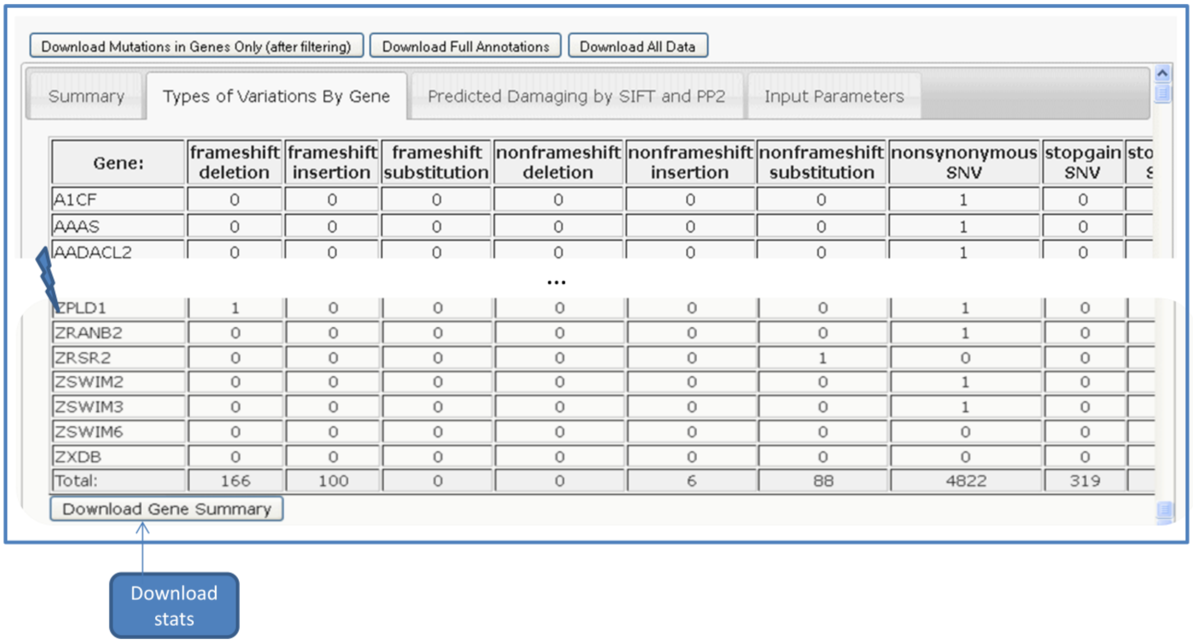
Download Mutations in Genes Only (after filtering)
Download Full Annotations
Download All Data
Summary
Types of Variations By Gene
Predicted Damaging by SIFT and PP2
Input Parameters

| Description | Image |
|---|---|
| Plot densit<br>Click here plots. | |
| Plot density using Circos<br>Click here to view and rearrange circos plots. | |
| Plot dens<br>Click he ots. | |
| Plot density using Circos<br>Click here to view and rearrange circos plots. | |

Tabs in Detail:

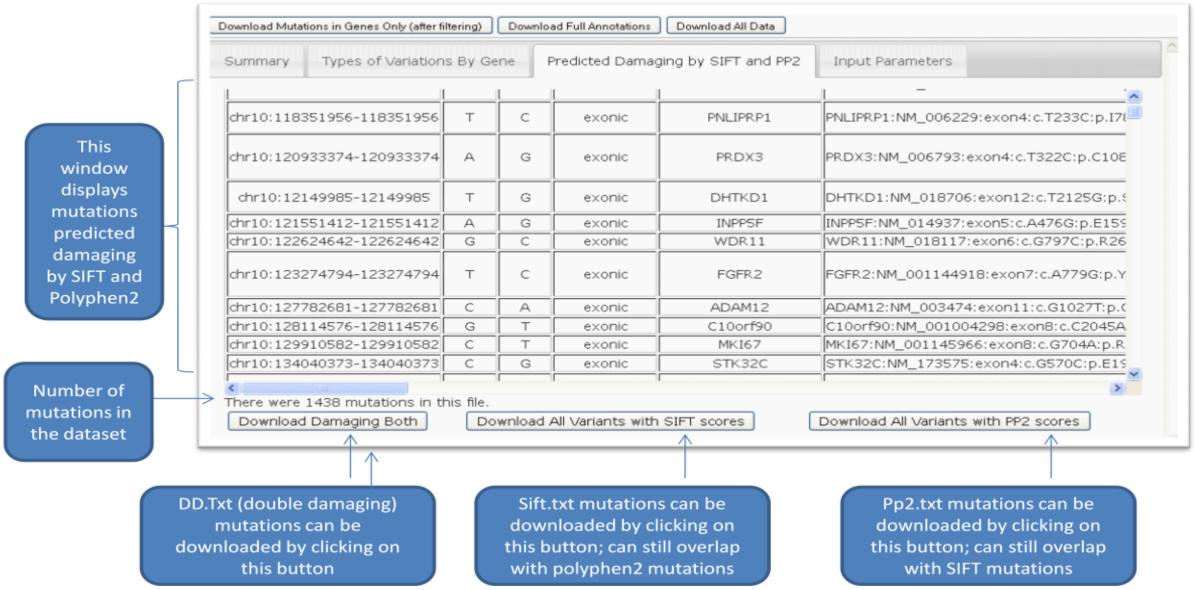
**Figure 32. Variation Types By Gene Tab**

This window shows a snippet of counts of mutation by type by gene. At the bottom of this tab, there is a running tally of the types of mutations types in the input dataset. These results contain all mutations by gene and mutations lost by population subtraction or gene filtering has not yet occurred.


Download Mutations in Genes Only (after filtering)
Download Full Annotations
Download All Data
Summary
Types of Variations By Gene
Predicted Damaging by SIFT and PP2
Input Parameters
Download Gene Summary
Download stats

| Gene: | frameshift deletion | frameshift insertion | frameshift substitution | nonframeshift deletion | nonframeshift insertion | nonframeshift substitution | nonsynonymous SNV | stopgain SNV | sto |
|---|---|---|---|---|---|---|---|---|---|
| A1CF | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | |
| AAAS | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | |
| AADACL2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | |
| | | | | ... | | | | | |
| ZPLD1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | |
| ZRANB2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | |
| ZRSR2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | |
| ZSWIM2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | |
| ZSWIM3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | |
| ZSWIM6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| ZXDB | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Total: | 166 | 100 | 0 | 0 | 6 | 88 | 4822 | 319 | |

**Figure 33. Damaging Mutations Tab**

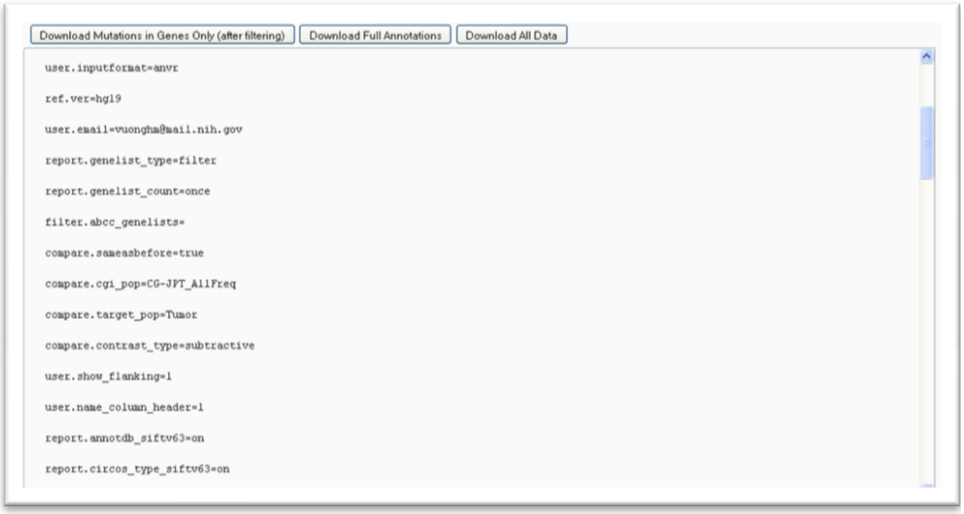

Download Mutations in Genes Only (after filtering)
Download Full Annotations
Download All Data
Summary
Types of Variations By Gene
Predicted Damaging by SIFT and PP2
Input Parameters
Download Damaging Both
Download All Variants with SIFT scores
Download All Variants with PP2 scores
This window displays mutations predicted damaging by SIFT and Polyphen2
Number of mutations in the dataset
DD.Txt (double damaging) mutations can be downloaded by clicking on this button
Sift.txt mutations can be downloaded by clicking on this button; can still overlap with polyphen2 mutations
Pp2.txt mutations can be downloaded by clicking on this button; can still overlap with SIFT mutations

| | | | | | |
|---|---|---|---|---|---|
| chr10:118351956-118351956 | T | C | exonic | PNLIPRP1 | PNLIPRP1:NM_006229:exon4:c.T233C:p.I7 |
| chr10:120933374-120933374 | A | G | exonic | PRDX3 | PRDX3:NM_006793:exon4:c.T322C:p.C10 |
| chr10:12149985-12149985 | T | G | exonic | DHTKD1 | DHTKD1:NM_018706:exon12:c.T2125G:p. |
| chr10:121551412-121551412 | A | G | exonic | INPP5F | INPP5F:NM_014937:exon5:c.A476G:p.E15 |
| chr10:122624642-122624642 | G | C | exonic | WDR11 | WDR11:NM_018117:exon6:c.G797C:p.R26 |
| chr10:123274794-123274794 | T | C | exonic | FGFR2 | FGFR2:NM_001144918:exon7:c.A779G:p.Y |
| chr10:127782681-127782681 | C | A | exonic | ADAM12 | ADAM12:NM_003474:exon11:c.G1027T:p. |
| chr10:128114576-128114576 | G | T | exonic | C10orf90 | C10orf90:NM_001004298:exon8:c.C2045A |
| chr10:129910582-129910582 | C | T | exonic | MKI67 | MKI67:NM_001145966:exon8:c.G704A:p.R |
| chr10:134040373-134040373 | C | G | exonic | STK32C | STK32C:NM_173575:exon4:c.G570C:p.E1 |

There were 1438 mutations in this file.

The Parameters tab will tell you all the information needed to recreate the analysis you just performed, including the databases run, database version, AVIA versions and the other parameters selected on the input page.  Keep this with your analysis.

**Figure 34.  Input Parameters Tab**



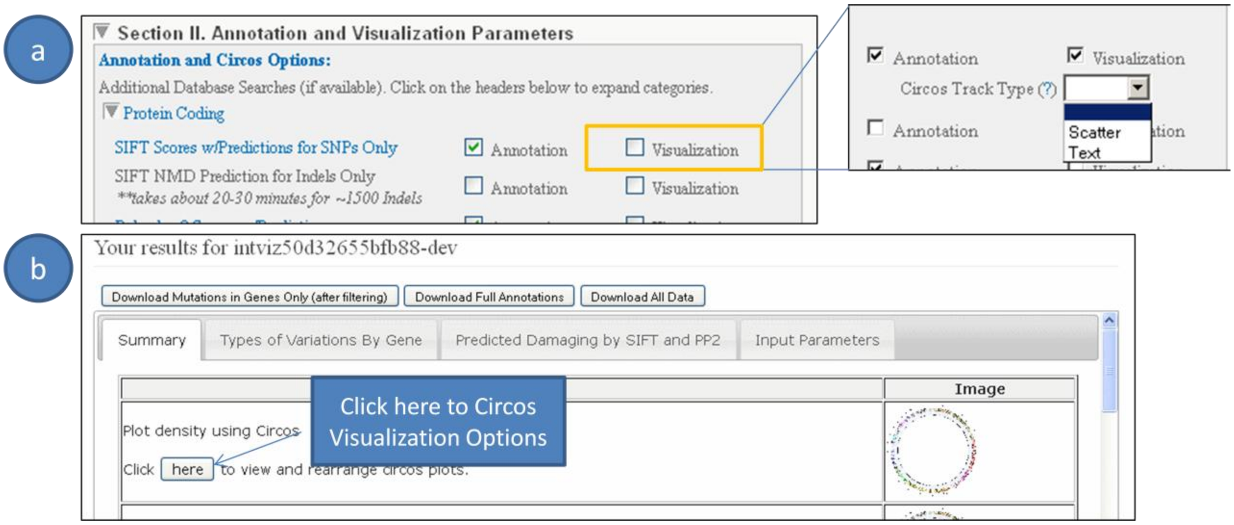
Download Mutations in Genes Only (after filtering)
Download Full Annotations
Download All Data

```
user.inputformat=anvr

ref.ver=hg19

user.email=vuonghm@mail.nih.gov

report.genelist_type=filter

report.genelist_count=once

filter.abcc_genelists=

compare.sameasbefore=true

compare.cgi_pop=CG-JPT_AllFreq

compare.target_pop=Tumor

compare.contrast_type=subtractive

user.show_flanking=1

user.name_column_header=1

report.annotdb_siftv63=on

report.circos_type_siftv63=on
```

## 2) Circos Visualization: Viewing and rearranging plots

Based on user request, AVIA generates Circos Plots of the user's input variation data.  A unique feature in AVIA is that users can rearrange multiple single plots generated by the AVIA pipeline into a multi-Circos plot without having to resubmit a request.  Recall from Workflows Section of the Tutorial, there were two options for each database in the Annotation and Visualizations Section; one for "Annotation" and one for "Visualization" as seen in Figure 35a.

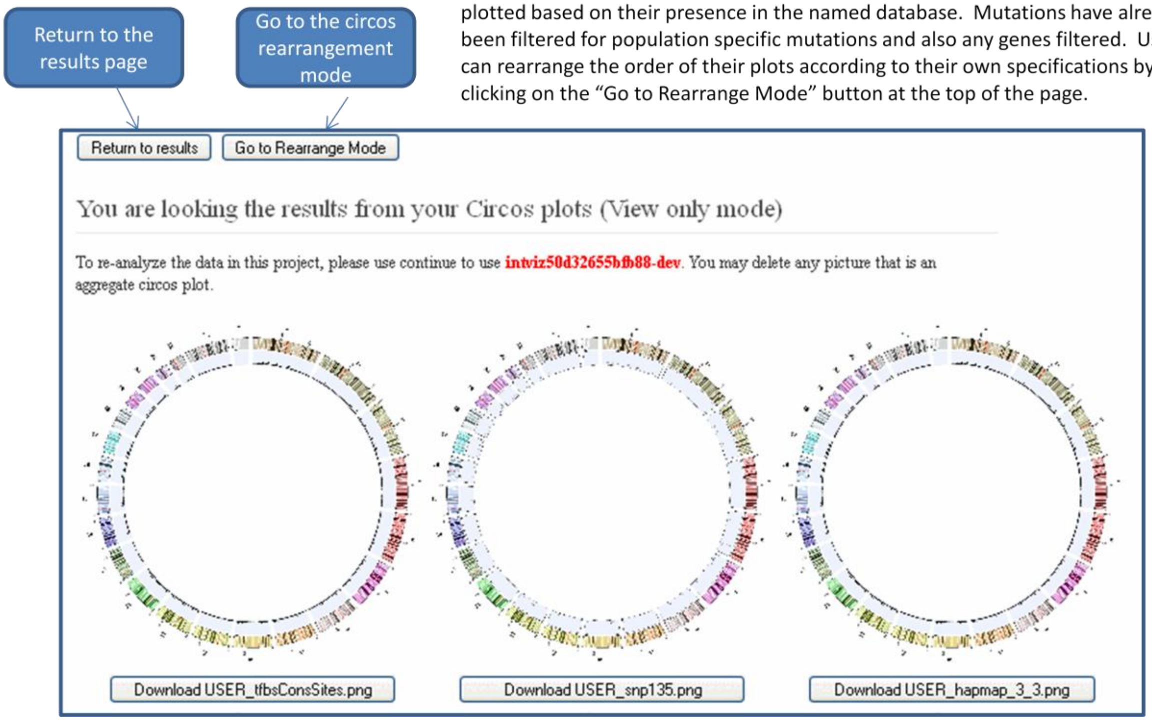
**Figure 35.  Circos Visualization Checkbox(es)**



a
Section II. Annotation and Visualization Parameters
Annotation and Circos Options:
Additional Database Searches (if available). Click on the headers below to expand categories.
Protein Coding
SIFT Scores w/Predictions for SNPs Only
☑ Annotation
☐ Visualization
SIFT NMD Prediction for Indels Only
**takes about 20-30 minutes for ~1500 Indels
☐ Annotation
☐ Visualization
☑ Annotation
☑ Visualization
Circos Track Type (?)
☐ Annotation
Scatter
Text
b
Your results for intviz50d32655bfb88-dev
Download Mutations in Genes Only (after filtering)
Download Full Annotations
Download All Data
Summary
Types of Variations By Gene
Predicted Damaging by SIFT and PP2
Input Parameters
Image
Plot density using Circos
Click here to view and rearrange circos plots.
Click here to Circos Visualization Options

Once the Visualization box was selected, an additional option to select the “Circos Track type” appeared, inset Figure 35a. If you select this option, you will be able to continue to the visualization portion described next. If you selected multiple databases for visualization, then you will have Circos images on your Results landing page to click on, as seen in Figure 35b. By clicking on any of the “Click here to view and rearrange Circos Plots” buttons on your results landing page, you will be redirected to the page with all of your Circos plots.

**Figure 36. Circos Visualization Page**


Return to the results page
Go to the circos rearrangement mode
Return to results
Go to Rearrange Mode
Download USER_tfbsConsSites.png
Download USER_snp135.png
Download USER_hapmap_3_3.png

By default, circos plots are generated one database per plot. Genic mutations are plotted based on their presence in the named database. Mutations have already been filtered for population specific mutations and also any genes filtered. Users can rearrange the order of their plots according to their own specifications by clicking on the “Go to Rearrange Mode” button at the top of the page.

You are looking the results from your Circos plots (View only mode)

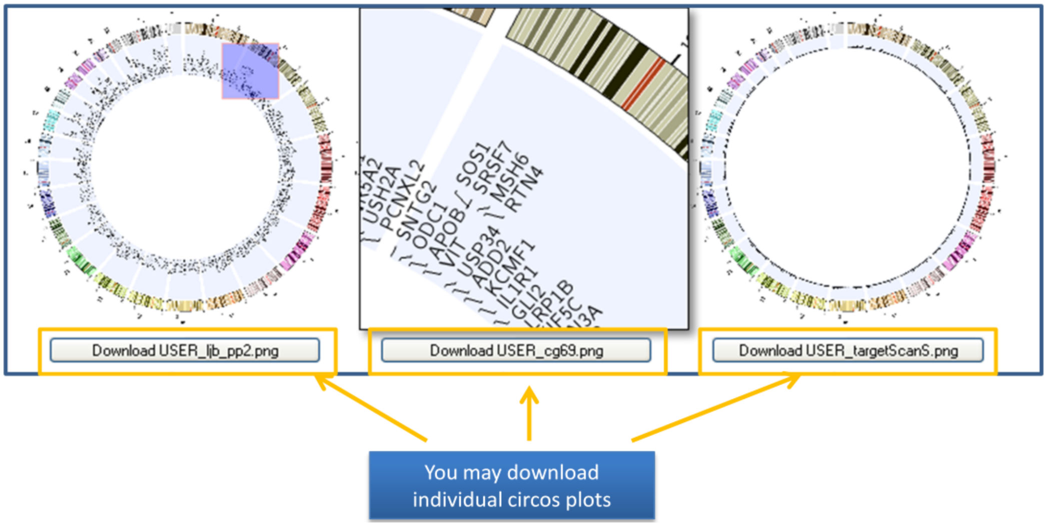
To re-analyze the data in this project, please use continue to use intviz50d32655bfb88-dev. You may delete any picture that is an aggregate circos plot.

The following pages are Circos specific pages and may not be of interest if you only have one plot. Figure 36 (above) is the first Circos visualization page with your user data.

On the page shown in Figure 37, you can view all of your plots. You can zoom in and move around the Circos plots to take a closer look (Figure 37, inset). You also have the option to download individual plots by clicking on the download button below the Circos plots. (orange boxes in Figure 37). This page is strictly for viewing plots. In order to combine multiple plots for publication or general viewing, you can navigate to the “Rearrange Mode” on the top of the page.

**Figure 37. Circos Visualization Page Zoom Option**

On this page, users may also zoom into their circos plots. Hovering over the plot will magnify the region in the light gray box.

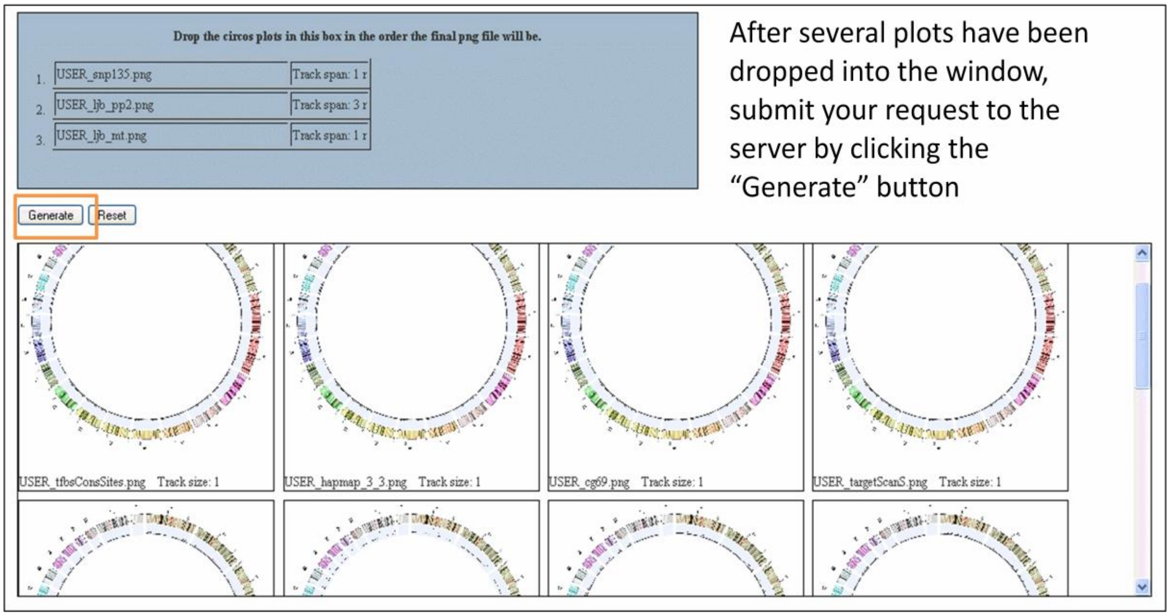


You may download individual circos plots

By clicking on the “Rearrange Mode” on the top of your Circos page, you will be directed to the rearrangement mode. This is where you can make Circos plots with more than one ring.

**Figure 38. Circos Rearrange Page.**

Circos plots can be dragged and dropped into the grey window. Plots will disappear from the lower window when dropped in the grey window. The order in the window will determine the rings of the circos plot.



Drop the circos plots in this box in the order the final png file will be.

1. USER_snp135.png Track span: 1 r
2. USER_ljb_pp2.png Track span: 3 r
3. USER_ljb_mt.png Track span: 1 r

Generate
Reset
USER_tfbsConsSites.png Track size: 1
USER_hapmap_3_3.png Track size: 1
USER_cg69.png Track size: 1
USER_targetScanS.png Track size: 1

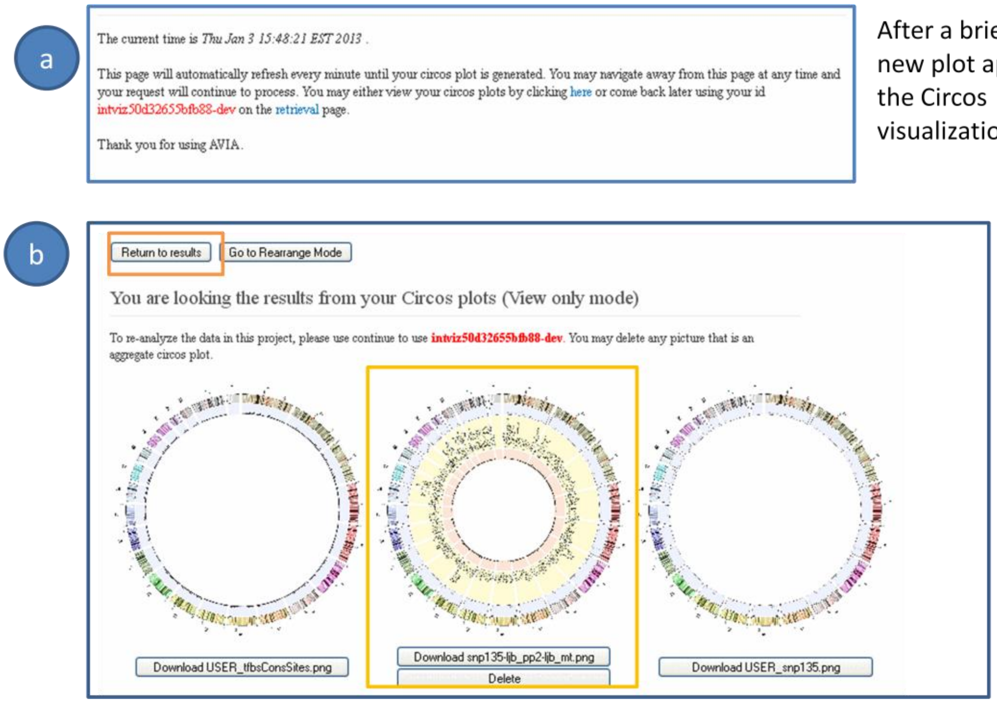
After several plots have been dropped into the window, submit your request to the server by clicking the “Generate” button

Only single plots with single rings will be displayed here. Each of the plots on the bottom panel can be dragged to the light gray box. The name and track span will appear in the table in the gray box if it has been successfully added. Each row in the table represents a ring within the Circos Plots in the order that it appears in the table. In the example shown in Figure 38, we plotted the data separately using 3 databases: snp135, ljb_pp2 and ljb_mt. We dragged the individual Circos Plots to the gray box so that we will generate a new Circos plot with three rings representing each of the three databases added in the order added to the gray box. You may then click on the "Generate" button, shown in the orange box, to send to AVIA to redraw your picture.

After a brief wait (Figure 39a), a new Circos plot with your tracks inserted in the order requested as shown in Figure 39b, large yellow box. You may do this multiple times.

You can return to your results page to download all of your data by clicking on the "Return to results" (shown in the orange box in Figure 39b) button at the top left corner of the navigation pane. From that page, there are several buttons to choose which data you wish to download, including your newly created Circos plots.

**Figure 39. Circos Processing and New Plot Page**



The current time is *Thu Jan 3 15:48:21 EST 2013* .

This page will automatically refresh every minute until your circos plot is generated. You may navigate away from this page at any time and your request will continue to process. You may either view your circos plots by clicking here or come back later using your id intviz50d32655bfb88-dev on the retrieval page.

Thank you for using AVIA.

You are looking the results from your Circos plots (View only mode)

To re-analyze the data in this project, please use continue to use **intviz50d32655bfb88-dev**. You may delete any picture that is an aggregate circos plot.

After a brief wait, a new plot appears in the Circos visualization page.

## 3) Downloading Results Archive

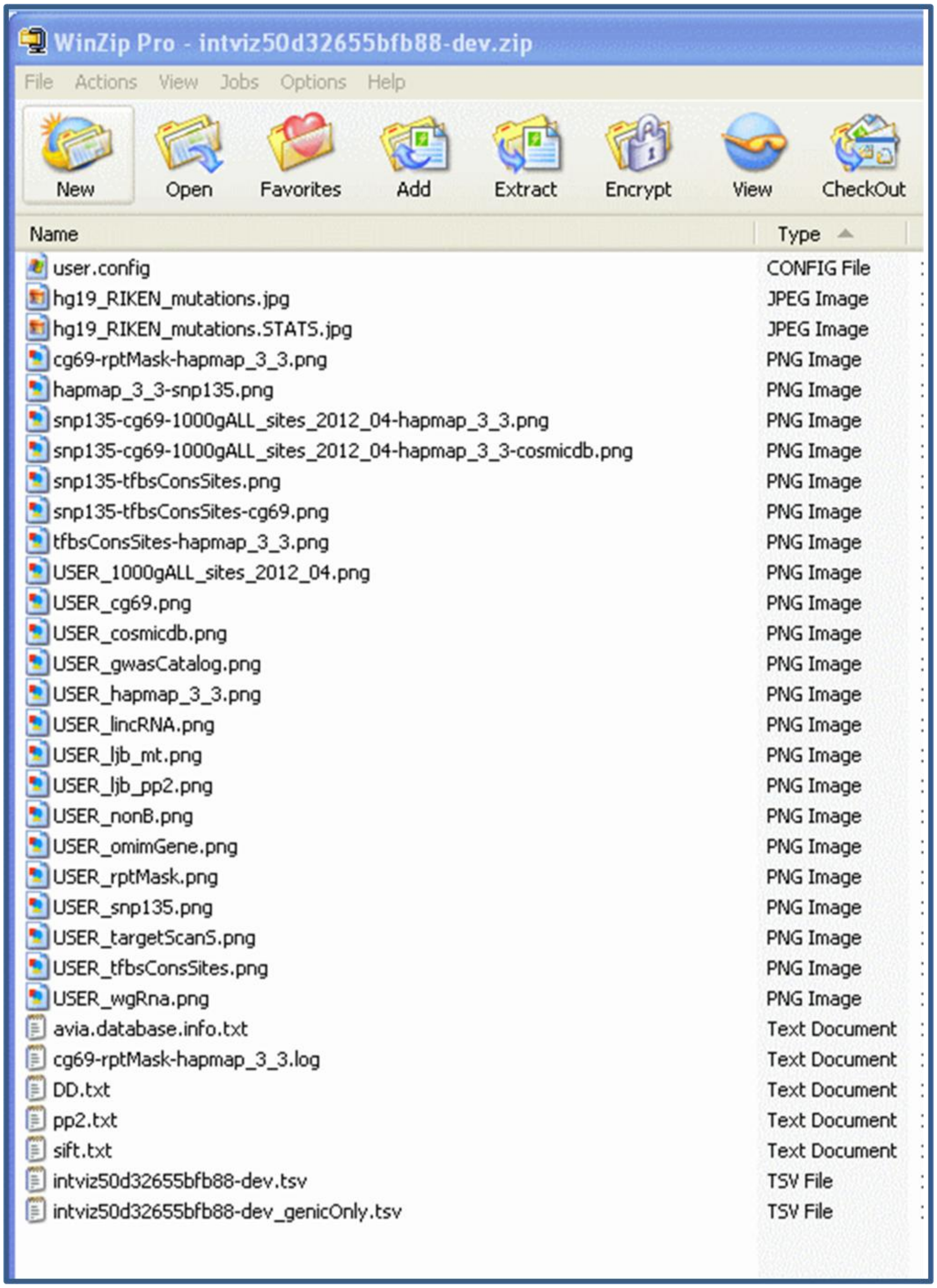
We have shown many places where you can download individual files. On the main results page, there is an option to "Download All Data" for your entire request. Below we describe in detail the contents included in the archive and how to view the data.

The contents of this zipped file can be viewed with any zip program such as WinZip (http://www.winzip.com/win/en/index.htm) or 7-Zip (http://www.7-zip.org/). Your archive may vary in number of files and different file names; however many of the files have standardized AVIA names. Filenames followed by asterisks (*) indicates that it is a name derived based on the user's original input filename. In the example shown in Figure 38, the input of our variants file was 'hg19_RIKEN_mutations.txt'. Some of the filenames may have derived from the user's original filename, which are in the light pink font. Other filenames are fixed or dependent on the database used.

- **user.config** : shows your inputs to AVIA, including your annotation and Circos options
- ***jpg files:** shows statistics in the user's variant data
  1. Venn diagram of damaging mutations according to sift and polyphen2 and snps found in dbSNP **(ex. "**hg19_RIKEN_mutations**.jpg")**
  2. Pie chart of exonic mutation types (ex. "hg19_RIKEN_mutations**.STATS.jpg**")
- ***png files :** show user's Circos plots using the name of the database plotted.
  1. Any files beginning with "USER" are the single Circos plots. e.g USER_cg69.png
  2. All others are the names of the multi-level plot and the names reflect the order in which the rings appear.
     eg. Cg69-rptMask-hapMap_3_3.png indicates a 3 ring Circos plot with cg69 mutations, rptMasker mutations and hapMap mutations plotted.
- **avia.database.info.txt :** file containing database information used for your input, including the build and other significant information.
- **DD.txt:** file that contains all mutations found damaging by both SIFT and Polyphen2.
- **pp2.txt:** file containing mutations that were damaging by Polyphen2.
- **sift.txt:** file containing mutations that were found damaging by SIFT.
- ***tsv files:** tab delimited files containing all your annotation for your input file using your unique AVIA identifier (*_genicOnly.tsv contains only genic mutations)

An example archive is shown in Figure 40. Your archive may more or less files than this depending on your input selections. In the user.config file, it will show which version of AVIA that was used in analysis.

**Figure 40. AVIA Archive Contents**



| Name | Type |
|---|---|
| user.config | CONFIG File |
| hg19_RIKEN_mutations.jpg | JPEG Image |
| hg19_RIKEN_mutations.STATS.jpg | JPEG Image |
| cg69-rptMask-hapmap_3_3.png | PNG Image |
| hapmap_3_3-snp135.png | PNG Image |
| snp135-cg69-1000gALL_sites_2012_04-hapmap_3_3.png | PNG Image |
| snp135-cg69-1000gALL_sites_2012_04-hapmap_3_3-cosmicdb.png | PNG Image |
| snp135-tfbsConsSites.png | PNG Image |
| snp135-tfbsConsSites-cg69.png | PNG Image |
| tfbsConsSites-hapmap_3_3.png | PNG Image |
| USER_1000gALL_sites_2012_04.png | PNG Image |
| USER_cg69.png | PNG Image |
| USER_cosmicdb.png | PNG Image |
| USER_gwasCatalog.png | PNG Image |
| USER_hapmap_3_3.png | PNG Image |
| USER_lincRNA.png | PNG Image |
| USER_ljb_mt.png | PNG Image |
| USER_ljb_pp2.png | PNG Image |
| USER_nonB.png | PNG Image |
| USER_omimGene.png | PNG Image |
| USER_rptMask.png | PNG Image |
| USER_snp135.png | PNG Image |
| USER_targetScanS.png | PNG Image |
| USER_tfbsConsSites.png | PNG Image |
| USER_wgRna.png | PNG Image |
| avia.database.info.txt | Text Document |
| cg69-rptMask-hapmap_3_3.log | Text Document |
| DD.txt | Text Document |
| pp2.txt | Text Document |
| sift.txt | Text Document |
| intviz50d32655bfb88-dev.tsv | TSV File |
| intviz50d32655bfb88-dev_genicOnly.tsv | TSV File |

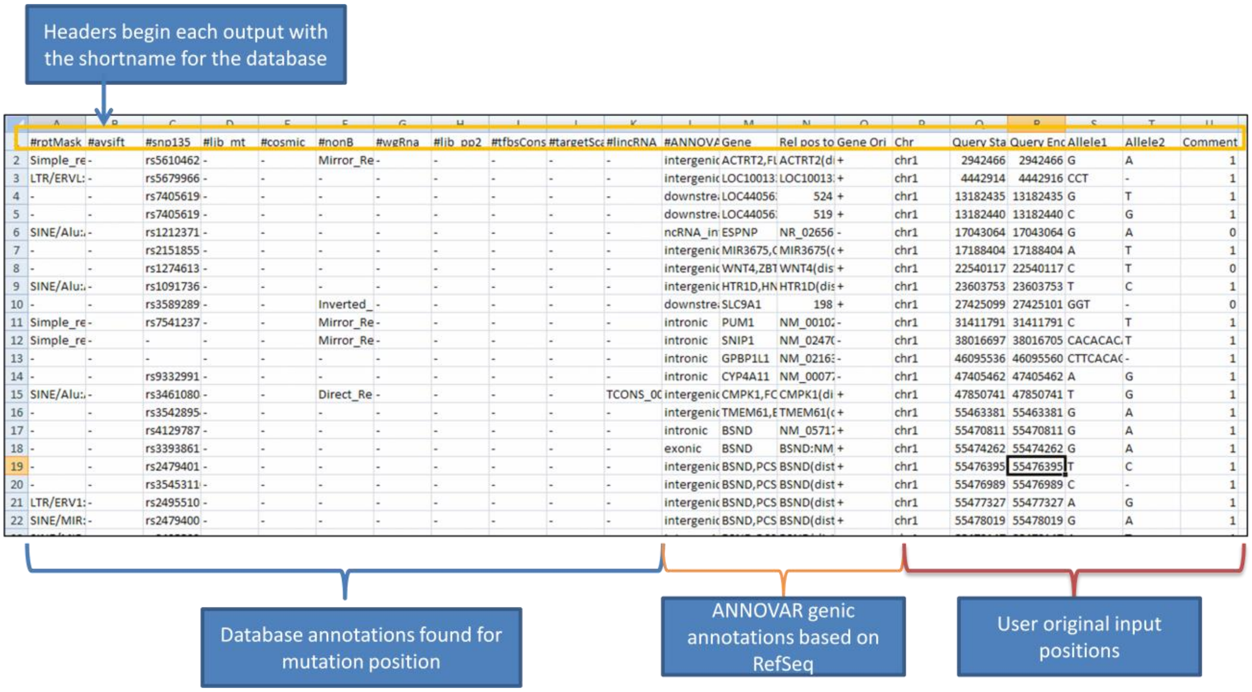
## B) Viewing Results on Local Computer

### 1) AVIA Output Annotation File (tsv)

The main output of AVIA is the annotation in a tab separated file which can be viewed in any text editor. In Figure 41, we show an example using Microsoft Excel®. The first line is always the header to describe what is in the column. Each annotation database is shown in the columns preceding genic annotation and user data. Each header is a brief descriptor of the contents of the column. Each line corresponds to variation data from the user's original input, each of the columns refers to any annotation associated with the database for a specific mutation. In Figure 41, we show that annotations were requested for rptMask, avsift, snp135, ljb_mt, cosmic, nonB, wgRna, ljb_pp2, tfbsConsSites, targetScan, and lincRNA. Whenever possible, we try to replace the database names with more meaningful descriptors. If you do not know to which databases the column belongs, you can refer to our "Databases">"View Databases" option under the User Navigation Menu match the "ANNOVAR Abbrv Name" to these headers.

DD.txt, pp2.txt, sift.txt, and *genicOnly.tsv are fixed filenames and are subsets of this file and are best viewed using Excel®. Other text files are best viewed using text editors such as Notepad ® ,Wordpad ® or Microsoft Word ®.

**Figure 41. AVIA Annotation Output File**



Headers begin each output with the shortname for the database

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | #rptMask | #avsift | #snp135 | #ljb_mt | #cosmic | #nonB | #wgRna | #ljb_pp2 | #tfbsCons | #targetSca | #lincRNA | #ANNOVA | Gene | Rel pos to | Gene Ori | Chr | Query Sta | Query Enc | Allele1 | Allele2 | Comment |
| 2 | Simple_re | - | rs5610462 | - | - | Mirror_Re | - | - | - | - | - | intergenic | ACTRT2,FL | ACTRT2(di | + | chr1 | 2942466 | 2942466 | G | A | 1 |
| 3 | LTR/ERVL: | - | rs5679966 | - | - | - | - | - | - | - | - | intergenic | LOC10013: | LOC10013: | + | chr1 | 4442914 | 4442916 | CCT | - | 1 |
| 4 | - | - | rs7405619 | - | - | - | - | - | - | - | - | downstre: | LOC44056: | 524 | + | chr1 | 13182435 | 13182435 | G | T | 1 |
| 5 | - | - | rs7405619 | - | - | - | - | - | - | - | - | downstre: | LOC44056: | 519 | + | chr1 | 13182440 | 13182440 | C | G | 1 |
| 6 | SINE/Alu:: | - | rs1212371 | - | - | - | - | - | - | - | - | ncRNA_in | ESPNP | NR_02656 | - | chr1 | 17043064 | 17043064 | G | A | 0 |
| 7 | - | - | rs2151855 | - | - | - | - | - | - | - | - | intergenic | MIR3675,C | MIR3675(c | + | chr1 | 17188404 | 17188404 | A | T | 1 |
| 8 | - | - | rs1274613 | - | - | - | - | - | - | - | - | intergenic | WNT4,ZBT | WNT4(dis | + | chr1 | 22540117 | 22540117 | C | T | 0 |
| 9 | SINE/Alu:: | - | rs1091736 | - | - | - | - | - | - | - | - | intergenic | HTR1D,HN | HTR1D(dis | + | chr1 | 23603753 | 23603753 | T | C | 1 |
| 10 | - | - | rs3589289 | - | - | Inverted_ | - | - | - | - | - | downstre: | SLC9A1 | 198 | + | chr1 | 27425099 | 27425101 | GGT | - | 0 |
| 11 | Simple_re | - | rs7541237 | - | - | Mirror_Re | - | - | - | - | - | intronic | PUM1 | NM_00102 | - | chr1 | 31411791 | 31411791 | C | T | 1 |
| 12 | Simple_re | - | - | - | - | Mirror_Re | - | - | - | - | - | intronic | SNIP1 | NM_0247( | - | chr1 | 38016697 | 38016705 | CACACAC. | T | 1 |
| 13 | - | - | - | - | - | - | - | - | - | - | - | intronic | GPBP1L1 | NM_0216: | - | chr1 | 46095536 | 46095560 | CTTCACAC | - | 1 |
| 14 | - | - | rs9332991 | - | - | - | - | - | - | - | - | intronic | CYP4A11 | NM_0007: | - | chr1 | 47405462 | 47405462 | A | G | 1 |
| 15 | SINE/Alu:: | - | rs3461080 | - | - | Direct_Re | - | - | - | - | TCONS_0C | intergenic | CMPK1,FC | CMPK1(di | + | chr1 | 47850741 | 47850741 | T | G | 1 |
| 16 | - | - | rs3542895 | - | - | - | - | - | - | - | - | intergenic | TMEM61,E | TMEM61(c | + | chr1 | 55463381 | 55463381 | G | A | 1 |
| 17 | - | - | rs4129787 | - | - | - | - | - | - | - | - | intronic | BSND | NM_05717 | + | chr1 | 55470811 | 55470811 | G | A | 1 |
| 18 | - | - | rs3393861 | - | - | - | - | - | - | - | - | exonic | BSND | BSND:NM | + | chr1 | 55474262 | 55474262 | G | A | 1 |
| 19 | - | - | rs2479401 | - | - | - | - | - | - | - | - | intergenic | BSND,PCS | BSND(dist | + | chr1 | 55476395 | 55476395 | T | C | 1 |
| 20 | - | - | rs3545311 | - | - | - | - | - | - | - | - | intergenic | BSND,PCS | BSND(dist | + | chr1 | 55476989 | 55476989 | C | - | 1 |
| 21 | LTR/ERV1: | - | rs2495510 | - | - | - | - | - | - | - | - | intergenic | BSND,PCS | BSND(dist | + | chr1 | 55477327 | 55477327 | A | G | 1 |
| 22 | SINE/MIR: | - | rs2479400 | - | - | - | - | - | - | - | - | intergenic | BSND,PCS | BSND(dist | + | chr1 | 55478019 | 55478019 | G | A | 1 |

Database annotations found for mutation position

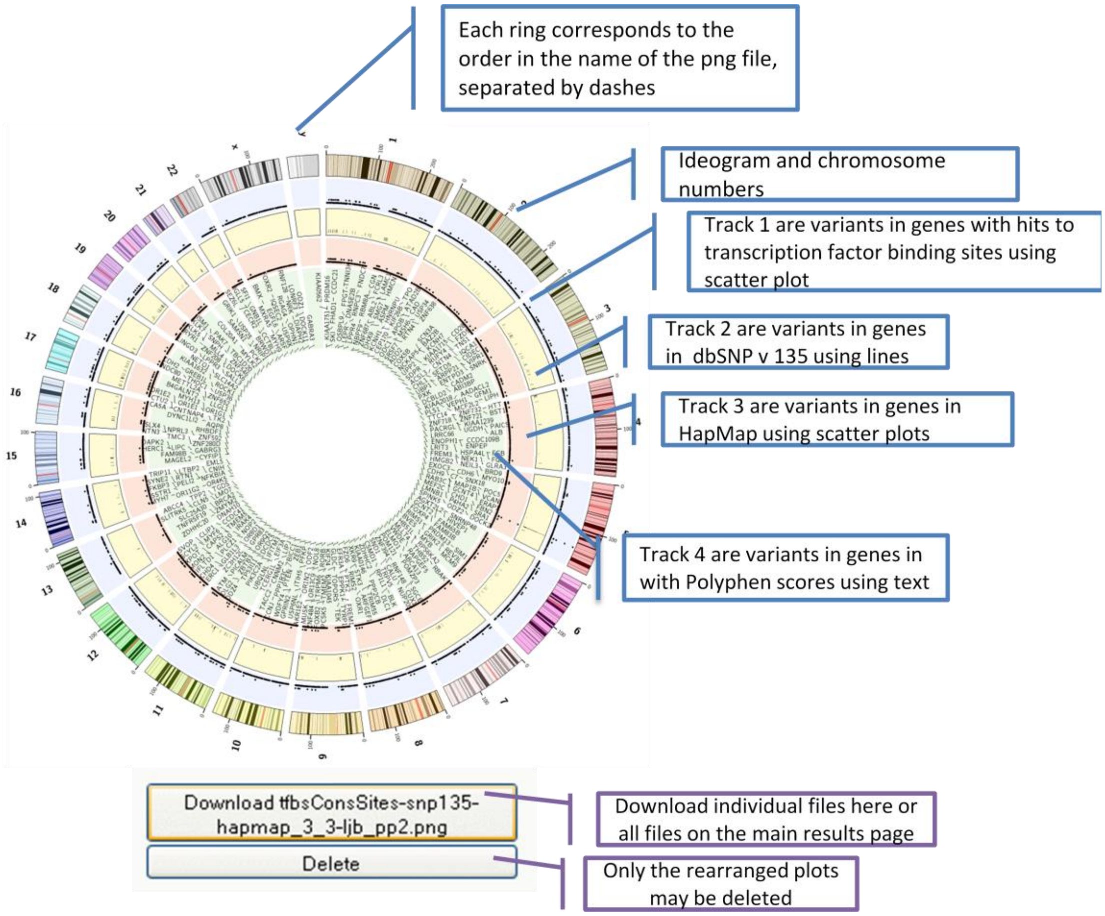
ANNOVAR genic annotations based on RefSeq

User original input positions

### 2) How to Read the Circos Plot Names

Below (Figure 42) is an illustration of how to interpret the Circos rings using the filename. In order to name figures uniquely and condense names, we had to abbreviate the names of the databases. Please refer to the "Databases">"View Database" in the User Navigation Panel along the top of any AVIA page.

**Figure 42. How to Read a AVIA-Circos Rearranged Plot**



Each ring corresponds to the order in the name of the png file, separated by dashes

Ideogram and chromosome numbers

Track 1 are variants in genes with hits to transcription factor binding sites using scatter plot

Track 2 are variants in genes in dbSNP v 135 using lines

Track 3 are variants in genes in HapMap using scatter plots

Track 4 are variants in genes in with Polyphen scores using text

Download individual files here or all files on the main results page

Only the rearranged plots may be deleted

### 3) Viewing graphics files

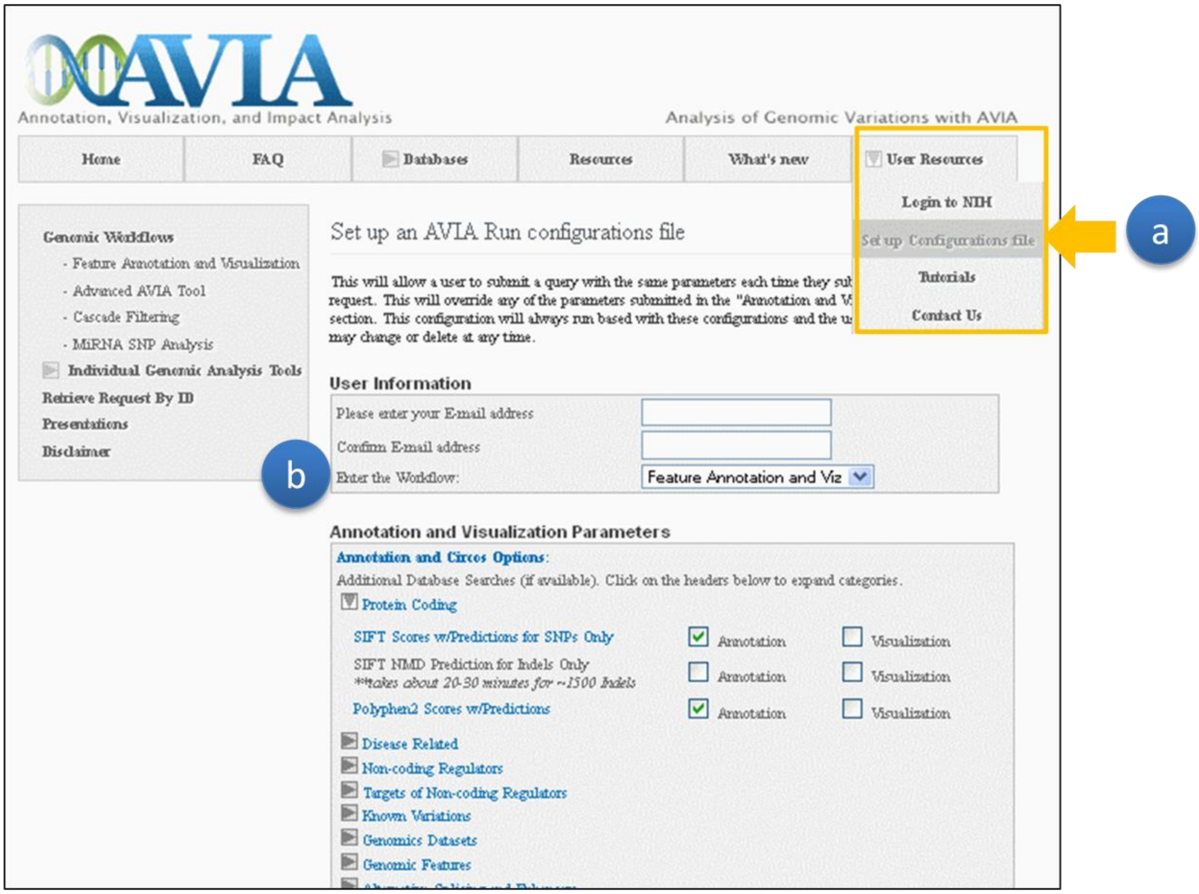
Each computer is different, but all of the graphics files (Venn Diagrams, Circos Plots) that end in ".png" or ".jpg" can be viewed with your computer's default image viewer. If you cannot open them, refer to your computer's user manual to find the appropriate program to view these types of files.

## VII) Reusable Configuration File

As you become more accustomed to the AVIA pipeline, you can set up a configuration file in which the minimal databases and features will be applied to each of the user's submitted runs, without having to click each of the desired database and features. Users will still need to fill out the "Input Data" section each time they upload, but based on the same email address used for the configuration set up, all of the queries and features will be applied with the new data file. This applies only to the “Annotation and Visualization” section of the web pages. On the individual submissions page, users may always add databases and features to the output. Users can also change their configuration at any time by following the directions on the page and confirm each submission by email. We currently do not accept batch loading of data files at this time.

To set up a configurations file, navigate to “User Resources” from the top navigation pane, and select “Set up a configurations file” as indicated by Figure 43a yellow arrow and box. You have two choices for the Workflow: Feature Annotation/Visualization and Cascade Filtering. Although these workflows are similar and have the same features, we found that not all users will want to filter their dataset each time so we have separated these into two distinct workflows shown in Figure 43b. The Cascade filtering workflow includes the Feature Annotation and Visualization. When you have completed the form, you will receive an email confirmation and once this link is activated by following the provided link, your configurations will be set up for every subsequent submission for the same workflow. If you wish to change your configurations, you may go to the same link, and change your information there. It will prompt you to delete your previous submission and will email you another confirmation link to activate your configurations file.

**Figure 43. Set up a Configurations File**



Set up an AVIA Run configurations file

This will allow a user to submit a query with the same parameters each time they sub request. This will override any of the parameters submitted in the "Annotation and V section. This configuration will always run based with these configurations and the u may change or delete at any time.

User Information

Please enter your E-mail address

Confirm E-mail address

Enter the Workflow: Feature Annotation and Viz

Annotation and Visualization Parameters

Annotation and Circos Options:

Additional Database Searches (if available). Click on the headers below to expand categories.

Protein Coding

| | | |
|---|---|---|
| SIFT Scores w/Predictions for SNPs Only | ☑ Annotation | □ Visualization |
| SIFT NMD Prediction for Indels Only<br>**takes about 20-30 minutes for ~1500 Indels | □ Annotation | □ Visualization |
| Polyphen2 Scores w/Predictions | ☑ Annotation | □ Visualization |

## VIII) References

1. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. Nat Methods **7**(4):248-249 (2010).
2. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: Circos: an information aesthetic for comparative genomics. *Genome Res* 2009, 19(9):1639-45.
3. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc. 2009*;4(7):1073-81.
4. Ng PC, Henikoff S. Predicting the Effects of Amino Acid Substitutions on Protein Function *Annu Rev Genomics Hum Genet.* 2006;7:61-80.
5. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003 Jul 1;31(13):3812-4.
6. Wang K., Le,M., Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010, 38(16):e164.